An Aggregate Second Order Continuum Model for Transient Production Planning

by

Matthew Wienke

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2015 by the
Graduate Supervisory Committee:

Dieter Armbruster, Chair
Donald Jones
Carl Gardner
Rodrigo Platte
Christian Ringhofer

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

Factory production is stochastic in nature with time varying input and output processes that are non-stationary stochastic processes. Hence, the principle quantities of interest are random variables. Typical modeling of such behavior involves numerical simulation and statistical analysis. A deterministic closure model leading to a second order model for the product density and product speed has previously been proposed. The resulting partial differential equations (PDE) are compared to discrete event simulations (DES) that simulate factory production as a time dependent M/M/1 queuing system. Three fundamental scenarios for the time dependent influx are studied: An instant step up/down of the mean arrival rate; an exponential step up/down of the mean arrival rate; and periodic variation of the mean arrival rate. It is shown that the second order model, in general, yields significant improvement over current first order models. Specifically, the agreement between the DES and the PDE for the step up and for periodic forcing that is not too rapid is very good. Adding diffusion to the PDE further improves the agreement. The analysis also points to fundamental open issues regarding the deterministic modeling of low signal-to-noise ratio for some stochastic processes and the possibility of resonance in deterministic models that is not present in the original stochastic process.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Ever since the emergence of industrial engineering and operations research as recognized disciplines the problems of planning and controlling production in the manufacturing industries has been a key application domain. Dating back to the beginning of the 20th century, there is an extensive body of literature in these disciplines. Included in this body is the work of Harris (1915) and Arrow (1958) which involved inventory models and Modigliani and Hohn (1955), and Holt (1960) that focused on discrete-time production planning. In fact, it is the now classical formulation of production planning due to the latter authors that is considered in this research.

In its simplest form this production planning paradigm describes a firm operating in a market where it faces external demand which it tries to meet by utilizing a limited set of production resources that have limited ability to generate output in a given time period. By *capacity* what is meant is that the limitation on the amount of output that can be produced in a given time interval by a production resource. However, this term is a colloquialism as the precise definition and determination of capacity turns out to be a challenging problem, as discussed by Elmaghraby (2011). Moreover, a complete and exhaustive definition of production planning in full generality is a complex and somewhat contentious task (Kempf *et al.* (2011)) and therefore, this document will remain less rigorous in its definitions of such topics.

For the purpose of this research, it is taken as a meaningful definition of capacity that which describes the production resource's ability to convert a specified mix of

inputs into a specified mix of outputs within a specified time frame. This definition gives rise to two closely related problems: the forward problem and the backward problem. The forward problem is to estimate within a desired level of accuracy the output trajectory over time obtained from a production resource given a specified input trajectory. The backward problem, on the other hand, is that of determining the necessary input pattern required to produce a desired output pattern also within a desired level of accuracy and over time. External demand we will assume to be a deterministic quantity that is known a priori. This simplification is not wholly unrealistic as it is common practice to schedule production activities based on available demand forecasts and this assumption corresponds to that practice. It will, though, be assumed that the actual behavior of the production resources is stochastic in nature. This stochasticity reflects various random influences such as unplanned machine breakdown and the natural variation in task production ties and material flows within the system. Both the forward problem and the backward problem treat the capacity of the system as a functional in the manner suggested by Hackman (2008) with its domain the set of all possible input trajectories and the range the set of all possible output trajectories (both over time) that the system is capable of producing.

The forward problem has been approached in several different ways in the literature. Very simple models, such as assuming that any input will be converted into output after a fixed time delay - widely assumed in the literature - are possible. Another approach is to develop a set of linear or nonlinear equations that represent the postulated behavior of the production system and solve these for the values of the output variables given the input quantities. However, one of the most widely used approaches is that of discrete event simulation (DES), in which a detailed simulation model of the production resource or facility is built and validated. This approach

will be discussed in a later chapter. A closely related approach is the use of detailed scheduling algorithms that develop detailed schedules for the loading of production resources over time, and thus describe their output.

The primary difficulty with the forward approach is that of the computational burden required to obtain a sufficiently accurate estimate of the output trajectory. This is mainly because the two aforementioned popular techniques, simulation and scheduling models, require increasing amounts of computational time as the number of resources and number of different tasks to be scheduled increases. The forward problem also suffers from the fact that it is descriptive in nature. Given a specified input pattern, it describes a predicted output pattern from which one may calculate desired performance measure estimates. It may also be possible to calculate other quantities that are not dependencies of either the input or output patterns, such as inventory levels over time if one has these two patterns in hand.

The backward problem, on the other hand, is inherently prescriptive in nature and production planning in its classical form, as first formulated byModigliani and Hohn (1955), addresses this problem. This approach seeks to determine an input pattern over time to a production resource or system that meets some external demand at a minimum cost. In this case, the formulation specifies the quantity of each type of output demanded by the market and attempts to compute the pattern of input over time that will meet this demand in some optimal or near optimal manner. The results of this model prescribe the amount of each different type of input to be released over time into the production facility along with the estimated output pattern produced by this input pattern.

The production planning function has as its aim the coordination over time of activities between different parts of the organization and external partners like vendors and customers. This coordination is generally carried out at discrete points in time in industrial practice. It involves the consideration of available information and the building of a plan of activities, specifically the release quantities, for a number of periods into the future, referred to as the *planning horizon*. This approach has been the dominant paradigm in the literature for decades (Johnson and Montgomery (1974), Vollmann *et al.* (2005)) and it renders a more aggregate approach in which one is not trying to compute input and output trajectories in continuous time, rather the total amounts of each associated with each of the discrete time periods in the planning horizon. This permits the use of simplified aggregate constraints to represent the behavior of the production resources, which simplifies the resulting optimization models.

An immediate issue that is raised by the classical production planning formulation described above is that of differing time scales. The discrete-time nature means that the planning activity takes place periodically. Yet, the actual factory for which these plans are generated operates in an essentially continuous manner. This temporal discrepancy introduces the possibility that the aggregate solution produced by the model for each time period (which could be days/weeks/months) may be impossible to implement on the shop floor due to detailed resource constraints (varying on the order of minutes/hours, etc) that have been ignored in the aggregation process. This can be somewhat mitigated by reducing the length of the planning horizons, however, adapting a too short planning horizon can, and in practice often does, bring about a new set of obstacles that are more severe than missing a demand target or beyond the scope of the model's influence. Thus, for a planning model to be able to produce even moderately accurate estimates of the input pattern required to produce

a desired output pattern and still maintain computational tractability requires that it incorporate some model of how decisions that are made at the planning level will affect the performance at the factory floor level during the planning horizon. Such a model is called an *anticipation function* (Schneeweiss (2003)).

Both theory and industrial practice suggest that the development of effective anticipation functions is a complex task for even simple production systems. A critical quantity for this purpose is the *cycle time* τ, the time elapsing between the work being released into the production system and its emergence as finished products that can be used to fulfill demand. Under steady-state conditions the cycle time is a nonlinear function of the resource utilization, which, in turn, is determined by the input trajectory determined by the planning models. When we take into account that production planning rarely treats systems that can be assumed to be in steady-state, and consider the number of resources and tasks involved in planning an even moderately sized production system, the fundamental nature of the challenges becomes apparent. The cycle time of any item moving through the production system is necessarily a random variable subject to some probability distribution that is a function of the resource utilization, the release trajectory up to that point in time, and the resources available to perform the necessary production tasks, among other factors. Here, the term *lead time* shall be used to denote the estimate of cycle time - which is usually the first moment - used in planning models.

This research will focus on the use of a continuous-time simulation model as an effective anticipation function. This model uses systems of coupled partial differential equations (PDEs) derived from transport equation models to estimate the output pattern of a production system from a given input pattern in continuous time. It

should be noted, however, that this approach is neither the definitive nor the most ubiquitous strategy employed in the field. This research will discuss, some more extensively than others, other prominent approaches to creating anticipation functions, highlighting their strengths and weaknesses, and illustrate how the transport equation models, specifically second-order hyperbolic models, are more effective in developing these functions.

Aside from the aforementioned transport equation models there are three significant models that are used as anticipation functions: Linear programming models with fixed lead times, DES and scheduling models, and the relatively recent clearing function models. It will discuss here in the introduction fixed lead time models as they are the simplest to understand and an exposition on them will serve to demonstrate the difficulties involved in developing anticipation functions as well as to introduce notation and concepts that will proliferate this document. The introduction will also introduce clearing functions since these models are an active area of research and run parallel to this research in many ways. Discussion of discrete event simulation (DES) and scheduling models will be more limited. A more extensive discussion of DES models is in a later chapter but only those aspects that pertain to model validation. Scheduling models will not be expounded upon since they contribute little, if anything, to this research approach–even though they are a very popular and effective stratagem in production planning.

1.1 Linear Programming (LP) Models with Fixed Exogenous Lead Times

For simplicity, production planning models for a single resource producing a single product is the starting point. Time will be divided into discrete periods $t = 1, \ldots, T$

that may not be of equal length. The goal is to determine the amount of material R_t that must be released to the resource during period t in order to meet a deterministic, known demand D_t in that period.

If a very short cycle time of the production resources being planned relative to the length of the planning periods is maintained one can achieve the simplest type of anticipation function. In this case, any material released for processing in period t is assumed to become available for use by the end of this period. Denoting the amount of finished goods available in inventory at the end of period t by I_t, the material balance equation describing the flows of material into and out of the inventory is given by

$$I_t = I_{t-1} + R_t - D_t. \tag{1.1}$$

Observe that one can dispense with the usual X_t variable denoting the amount of production in period t, because $X_t = R_t$ by assumption. Also, considerations of the work in progress (WIP) can be ignored since the material remains in the system for a very short duration due to the very short production resource cycle time. It is possible that the cycle time of the production resource may extend over more than one planning period. In the event, a common assumption is that the lead time is fixed and independent of the resource utilization, i.e. of the release quantities R_t. One encounters this approach commonly in both inventory theory (Zipkin (2000)) and mathematical programming approaches to production planning (Johnson and Montgomery (1974), Missbauer and Uzsoy (2011)), as well as the material requirements planning (MRP) approaches used widely in industry (Vollmann *et al.* (2005)). The stochastic equivalent of this model is a random lead time L with a time-stationary probability distribution that is independent of the order quantities such as the case

treated by Eppen and Martin (1988). In this case, the amount R_t released into the system during period t becomes available for use in period $t+L$. Denoting the output of the production system in period t by X_t, then the relationship $X_t = R_{t+L}$ and the system dynamics are now described by the relationship

$$I_t = I_{t-1} + X_t - D_t = I_{t-1} + R_{t-L} - D_t. \tag{1.2}$$

It is common in both the literature and in practice to assume that the fixed lead time L corresponds to an integer number of planning periods. As it stands, this model assumes that there is no limit on the amount of output the system can produce in the given lead time. Most optimization models of capacitated production systems will limit the total output of the system in a given period by imposing an aggregate capacity constraint of the form $X_t \leq C_t$, where C_t denotes the maximum possible output of the production resource in a given period. Consider now, for the purposes of exposition, that each unit produced requires exactly one time unit of the resource, and that the resource capacity is expressed in terms of time units available per planning period. Then, under this representation, the production resource can produce any amount of output up to C_t units in any period, and no material remains in the system for more than L time periods, regardless of the WIP level. Thus at the end of period t, the production system in this model will have a WIP level of

$$W_t = \sum_{n=t-L+1}^{t} R_n \; - \; \sum_{n=t+1}^{t+L} X_n \tag{1.3}$$

units of product. Note that only R_{t-L} units of WIP are actually available for the resource to convert into output in period t.

The issue of multiple time scales raises it head again when examining the aggregate capacity constraint described above, even though the production resource cycle times are short relative to the planning periods. This is due to the timing of event on the shop floor that are lost in the aggregation. As an example, suppose that the planning model recommends that $R_t = 200$ units be released into the system during period t, where the length of the time period is one week. If the capacity of the resource is $C_t = 200$ units per week, this will appear to be a feasible solution. However, if events conspire to have all 250 units arrive in the middle of the week, it very well may not be possible for the production resource to service all this material within the allotted period t. Subject to these limitations, one can therefore, describe the behavior of the production resource in a given period t with the set of constraints

$$X_t = R_{t-L}, \tag{1.4}$$

$$I_t = I_{t-L} + X_t - D_t, \tag{1.5}$$

$$X_t \leq C_t. \tag{1.6}$$

With these constraints in mind the conventional view of production capacity used in MRP and most mathematical programming models results in a linear program of the following form

$$\min \sum_{t=1}^{T} (h_t I_t + c_t R_t) \tag{1.7}$$

where h_t and c_t are the per unit inventory cost and the per unit production cost, respectively. This objective is subject to

$$I_t = I_{t-L} + X_t - D_t, \tag{1.8}$$

$$X_{t-\tau} \leq C_t, \tag{1.9}$$

$$R_t, I_t \geq 0 \tag{1.10}$$

for $t = 1, \ldots, T$.

The deficiency of this model is that it assumes that WIP will not accumulate in the system over time; the releases in the period $t - L$ constitute the entire WIP available to the resource for service in period t. these releases are implicitly constrained not to exceed the capacity, so the system is always able to process the entirety of its available WIP in a single period. The remainder of the WIP, given by

$$\sum_{n=t-\tau+2}^{t} R_n, \tag{1.11}$$

has no effect on the cycle time of the resource, which is always equal to the pre specified parameter τ. As far as this model of production capacity goes, it is completely unrelated to the capacity C_t of the resource in a given period. All the lead time L serves to do is delay the arrival of work at the resource after its release into the system. It does not describe the behavior of the resource itself, which is assumed in the capacity constraint to be able to service any amount of product up to the capacity limit C_t in a given period.

Now what has been chosen is a simple objective function, that of minimizing the sum of production and inventory holding costs over the planning horizon, and it must be pointed out that most LP models encountered in practice will involve additional constraints specific to the application domain under study as discussed by Hackman and Leachman (1989). Yet absent these constraints, the above model does represent the essentials of inventory balance between periods and aggregate capacity within periods. Moreover, if it can be assured that there are nonnegative inventory levels at the boundaries between periods then it is guaranteed that there will be nonnegative inventory throughout the period. This is because all the rates are uniformly distributed over a planning period. Of course, more elaborate objective functions are

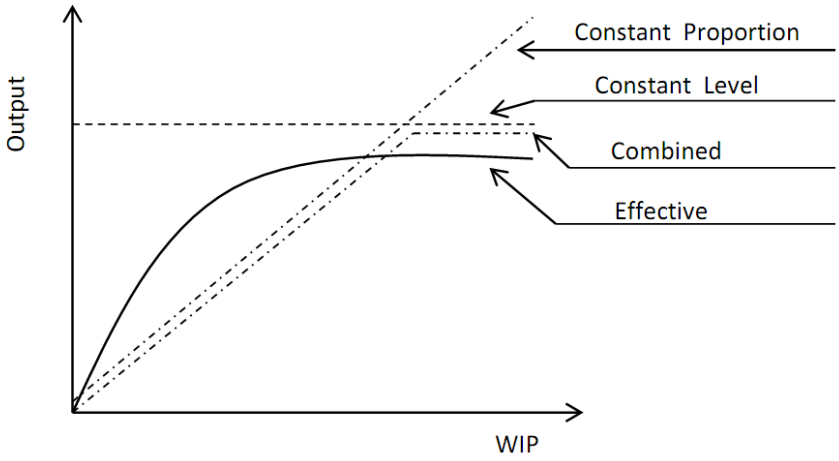
possible, but the above objective is sufficient to represent the production capacity and system dynamics of this fixed lead time approach.

1.2 LP Models and Clearing Functions

The ability of LP models to capture the nonlinear relationships between workload and cycle time for production systems that are governed by queueing is questionable. This is especially true in the event that resources are heavily utilized of when the utilization can vary significantly over time. The use of nonlinear clearing functions representing the expected output of a production resource as a function of some measure of the workload, usually the amount of WIP awaiting service, have been proposed in recent years and shows considerable promise. Only a brief overview of this topic will be given here. For a more complete review, the interested reader is referred to Missbauer and Uzsoy (2011).

In their most general form, clearing functions represent the relationship between the expected output of a production resource in a given planning period and some measure of the expected workload in that period. Several examples of clearing functions in the literature are illustrated in Fig 1.1. The constant level function represents the maximum allowable level of production. It does not have any lead time constraint and assumes instantaneous production independent of the WIP level W_t. Graves (1986) proposed a clearing function in the form of $X_t = \alpha W_t$, where the output X_t at time t is considered a linear function of the WIP. This *constant proportion* function assumes that a fixed lead time of $1/\alpha$ can be maintained at all utilization levels. In this model, it is assumed that the production facility will be operated in the range that this fixed lead time assumption will hold. At high WIP levels, though, this func-

Figure 1.1: Examples of Clearing Functions. (Karmarkar (1989))



tion may yield infeasible output values. To rectify this, it needs to be bounded by a fixed output capacity less than or equal to the maximum production capacity. This is shown in the figure as the *combined* clearing function. Both Karmarkar (1989) and Srinivasan *et al.* (1988) have proposed concave, nondecreasing functions of the W_t as nonlinear clearing functions.

To motivate the use of nonlinear clearing functions instead of linear or combined clearing functions, consider a production resource that can be modeled on a $G/G/1$ queueing system in steady-state. The average number in the system, i.e., the expected WIP W, is given by Medhi (2002) as

$$W = \frac{c_a^2 + c_s^2}{2} \frac{\rho^2}{1-\rho} + \rho, \tag{1.12}$$

where c_a and c_s denote the coefficients of variation of the interarrival and service times, respectively, and ρ is the utilization of the server. Setting $c = (c_a^2 + c_s^2)/2$ and rearranging Equation (1.12), one obtains a quadratic in W whose positive root yields the desired ρ value. solving for ρ with $c > 1$ provides

$$\rho = \frac{\sqrt{(W+1)^2 + 4W(c^2-1)} - (W+1)}{2(c^2-1)}, \tag{1.13}$$

which has the desired concave form. When $0 \leq c < 1$, the other root of the quadratic will always give positive values for ρ. When $c = 1$ (as in an $M/M/1$ queue), Equation (1.12) reduces to $\rho = W/(1+W)$, which is again of the desired concave form. Then, for a fixed value of c the utilization, and therefore the throughput, increases with increasing WIP, but at a decreasing rate due to the variability in the service and arrival rates.

The use of clearing functions is not without its limitations, however. Output is limited by a function of the expected total load in these models. Additionally, they

do not consider the distribution of the arrival period of the work that is expected to contribute to the load in period t. In his paper, Missbauer (2002) addresses these issues and proposed an aggregate order release planning model that both determines the amount of work released in each planning period and can manage different load patterns without the need to include additional load balancing parameters.

In deriving clearing functions, one can use both analytical and empirical techniques. Analytical techniques involve the use of steady-state or transient queuing models whereas empirical techniques estimates from empirical data. Different authors have implemented somewhat different approaches. Agnew (1976) proposed a throughput function where the service rate is a function of the number in the queue and suggested using it in an optimal control policy context. Spearman (1991) derived a clearing function using closed queuing networks, conjecturing a relationship between mean cycle time and WIP, and taking one observation from simulation to specify congestion in the system. Asmundsson *et al.* (2006) formulated the clearing function as a relationship between the expected throughput of a resource in a planning period and the time-average WIP level at the resource during the period from empirical data.

Other authors (Karmarkar (1989), Missbauer (2002)) assumed that the clearing function depends on the expected workload, which they define as the sum of the work in progress available at the start of the period and the material released during the period. Zäpfel and Missbauer (1993) used a simulation model to estimate clearing functions based on the expected workload and observed discrepancies between planned and actual WIP in simulation. Missbauer (2009) showed that the clearing function depends on the work in progress at the beginning of each period due to the transient behavior of the system and suggested a transient clearing function. Selcuk

et al. (2008) derived transient clearing functions analytically using the Pollaczek-Khinchine mean value formula and Little's law.

The dearth of empirical data from industry–usually due to the inability to maintain controlled steady-state production for even one production level, let alone the multiple levels required to obtain a comprehensive depiction of the resource–has lead researchers to use DES models of production systems to obtain data on the workload and output in each planning period. A notable exception to this is Haeussler and Missbauer (2014), who fit clearing functions to data obtained from a manufacturer of digital storage media. They highlighted considerable differences between the simulated and empirical data, most notably in that the empirical data show significantly higher variability. This variability may be due to a variety of factors such as different labor allocation policies, failures, etc. that are not included in the simulation model.

Asmundsson *et al.* (2006) used a visual technique for fitting piecewise linear segments approximating a concave clearing function and reported favorable results. In a subsequent paper (Asmundsson *et al.* (2009)) they used linear regression to fit a concave clearing function that was then piecewise linearized by solving a nonlinear program. They found that the clearing functions thus obtained consistently overestimate the capacity of resource and suggested an empirical technique to correct this by ensuring that a specified percentage of the data points lies above the fitted function. Kacar and Uzsoy (2010) used different multiple regression models and found that the variable selection procedures do not seem to have significant effects on the quality of the production plans obtained by using LP models based on different fitted clearing functions. Albey *et al.* (2011) used nonlinear regression to fit their disaggregated clearing functions, obtaining locally optimal solutions using a standard convex non-

linear solver. The majority of this work has assessed the quality of the fits obtained from the empirical data based on conventional statistical measures such as correlation coefficients. Results indicate that this area requires further research; different clearing function forms yielding very high correlation coefficients can result in quite different performance when the production plans obtained by using them in an optimization model are implemented.

Finally, in recent work, Kacar (2012) took a different approach to the fitting of clearing functions where the parameters of the clearing functions are optimized to yield the best production plans. They assessed the quality of the clearing function fits by solving an optimization model using the clearing function estimated with that parameter set and then simulating the execution of the production plans in the system under study. Their approach was to start from the clearing function fit obtained from the linear regression and attempt to improve this through a simulation-optimization approach, specifically the simultaneous perturbation stochastic approximation (SPSA) algorithm (Spall (1998)), as well as several heuristics that attempt to reduce the required computational time. They found that significant improvements were possible over the clearing functions obtained by regression; in other words, when the clearing functions obtained by the SPSA approach are used to develop input trajectories that are then disaggregated and simulated, the clearing functions obtained by the SPSA approach have statistically better performance. In extensive computational experiments on a scaled-down semiconductor wafer fabrication process, Kacar (2012) showed that the production plans using clearing functions fit by regression yield statistically better production plans than the iterative approaches of Kim and Kim (2001) and Hung and Leachman (1996). However, the use of the simulation optimization to refine the parameter estimates used in the clearing functions yield substantial

improvements in realized performance over the clearing functions obtained from regression alone. The additional advantage of the clearing function approach is that the actual planning does not require any simulation, although extensive simulation runs are necessary to obtain the data needed to fit the clearing functions.

1.3 Goal of the Research

The aim of this research is to study a second order continuum model and show that it improves upon solving the forward production planning problem. Derived from kinetic theory, a system of hyperbolic partial differential equations (PDE) has been introduced that describe the evolution of conserved quantities (i.e. mass density and velocity) over time and then constructs an outflux profile for specific input patterns.

In subsequent chapters, the theoretical and numerical groundwork that is used to develop the model is discussed and a simulation methodology will be developed. Following this, in chapter 4, comes the introduction of the first continuum based models. These include the original ordinary differential equation (ODE) fluid models as well as the first order transportation models based on the concept of clearing functions. The advantages of these models are demonstrable, but they also suffer from certain incapabilities and inaccuracies.

The capstone of the work is found in chapter 5. After presenting the second order model developed by Armbruster *et al.* (2006a) its ability to represent the average behavior for timed dependent stochastic inputs is studied and compared with the DES and the first order model. The average over several thousand runs for a given scenario of the discrete event simulations is used as the baseline for comparison. It will be

shown that the second order model, in general, outperforms the first order models in three fundamental scenarios for the time dependent influx: An instant step up/step down of the arrival rate, an exponential step up/step down and periodic variation of the average arrival rate. Specifically when the overall utilization of the system is low or where the utilization is increasing, the second order model matches the DES nicely since the model allows for the formation of rarefaction waves. For high and decreasing utilization, the PDE performs less impressively due to the shock waves inherent in this model. The incorporation of a small diffusion source term can partially mitigate the effects of these waves, yet they cannot be eliminated completely as the dominant flow for the PDE must stay hyperbolic and a diffusion term with a magnitude great enough to eliminate the shocks would violate this condition on the model.

Additionally, the long term stationary output for periodic input patterns matches extremely well for periods on the order of two or more times the average cycle time. Since changes in the influx for many real world factories (e.g. semiconductor factories) have timescales similar to this, these results are less removed from reality than one might believe given the simplicity of the underlying assumptions.

Lastly, the comparative analysis of the DES and the PDEs points to fundamental challenges for a kinetic theory based PDE model: The scenario involving cyclic input patterns shows that i) hyperbolic PDEs are susceptible to resonances that are not shared by their DES counterparts; and ii) fast time varying influx rates will result in a low signal-to-noise ratio for the averages DES that cannot be resolved with a second order deterministic model such as a PDE. We end the document with a summary of the work as well as a brief discussion of future avenues of research and observations made during initial forays.

Chapter 2

MATHEMATICAL FUNDAMENTALS

2.1 Queuing Systems [1]

A queuing system can be described as customers arriving for service, waiting in a queue for service if it is not immediate, and upon entering service, leaving the system after service has been rendered. For the majority of queuing systems there are six basic characteristics of queuing processes that provide an adequate description of the system utilized: (1) arrival pattern of customers, (2) service pattern of servers, (3) number of service channels, (4) system capacity, (5) queue discipline, and (6) number of service stages. Figure 2.1 illustrates a schematic for a general $M/M/1$ queueing system.

Usually, the arrival and service processes are stochastic and so it is necessary to know the probability distribution of the interarrival times (the reciprocal being the *arrival* rate), that is, the times between successive arrivals and the service times which is the length of time that the customer is in service (the reciprocal of this being the *service* rate). One note must be made concerning the arrival and service patterns of

[1] Reference material for this section is drawn from Gross *et al.* (2008)

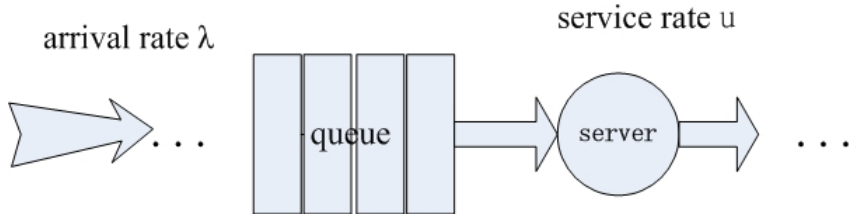


Figure 2.1: Schematic of a General $M/M/1$ Queueing System.

the customers: How does the pattern change with time? A *stationary* arrival pattern is one that does not change in time. This means that the probability distribution describing the input process is time-independent. Conversely, a time-dependent distribution is called *non-stationary*.

Even in the event of a high service rate (relative to the arrival rate) there is a non zero probability that customers will be waiting in the queue. Arrivals and departures from the queue happen at irregular intervals; consequently, the pattern for the length of the queue cannot be assumed without both the arrival and service processes being deterministic. It follows that the probability distribution for the queue length depends on both processes which are frequently, but not always, assumed to be independent.

Generally there are three system responses of interest for a given queuing system: (1) some measure of the waiting times for a customer; (2) a description of the manner in which customers may accumulate in the system; (3) a measure of the downtime (idle time) of the servers. Because stochastic elements comprise the majority of queuing systems, these measures are random and hence their probability distributions are highly desirable. However, it is quite difficult to determine the precise distributions for these measures for many systems *even in steady-state*. Luckily, the expected values often provide sufficient understanding of the system responses and can sometimes be more easily obtained.

The waiting times for customers come in two types, the time that a customer spends in the queue waiting until service and the total time that a customer is in the system (queue plus service). Correspondingly, there are two types of accumulation

measures also. Those are the number of customers in the queue at any given time and the total number of customers in the system at any given time. Finally, measuring the idle time of the servers can include either the time that any particular server is down (idle) or the time that the entire system is devoid of customers.

Let λ denote the mean arrival (influx) rate of the customers entering the system and μ as the mean service rate. Given that the average influx of customers is λ and that each one requires $1/\mu$ time units to be serviced (again, on average) a fundamental measure of effectiveness the system is the utilization $\rho \equiv \lambda/\mu$ which is the average number of customers in service in steady-state. Note that this measure is only defined for $\rho < 1$. Obviously when $\rho > 1$ the average influx rate exceeds the average service rate and the queue becomes unstable as time $\to \infty$. Less obvious is that the same occurs when $\rho = 1$. [2]

The probability distribution for the total number of customers in the system at time *t*, *N(t)*, is the sum of the number of customers in the queue, $N_q(t)$, and the number of customers currently in service, $N_s(t)$. Let $p_n \equiv Pr\{N = n\}$ in be the respective non-steady-state and steady-state probabilities for the number in the system. Two useful measures (expected-value) for a single server queuing system are the mean number in the system

$$W \equiv E\{N\} = \sum_{n=0}^{\infty} np_n, \tag{2.1}$$

and the mean number in the queue,

$$W_q \equiv E\{N_q\} = \sum_{n=2}^{\infty} (n-1)p_n. \tag{2.2}$$

[2]Unless *both* the arrival and service rates are deterministic and perfectly scheduled.

In the early 1960s, John Little [Little 1961] discovered a powerful relationship between the steady-state mean system sizes and the steady-state customer waiting times as follows. Let T_q be the time that a customer spends waiting in the queue prior to service and T_s the length of time that a customer spends in service. Then the total time, T, spent in the system for a customer is $T = T_q + T_s$. Since the arrival and service rates are stochastic, T, T_q, and T_s are all random variables with first moments τ, τ_q, and τ_s, respectively. Then Little's formulae are

$$W = \lambda\tau \tag{2.3}$$

and

$$W_q = \lambda\tau_q. \tag{2.4}$$

These results along with the fact that $E\{T\} = E\{T_q\} + E\{T_s\} \Leftrightarrow \tau = \tau_q + 1/\mu$ show that it is only necessary to find one of the four expected-value measures above for a queue in steady-state.

A simple yet relevant type of queue for which one can determine the steady-state probabilities is the $M/M/1$ queue as illustrated in Figure 2.1. The $M/M/1$ queue assumes exponential inter arrival times, a single server with exponential process times, no buffer cap, and a first-come first-served queue protocol. Because it deals with an exponential distribution, knowledge of the mean arrival and service rates is sufficient to completely describe these distributions. Furthermore, the memoryless property of the exponential distribution allows one to express the state of the system via one number, N, representing the number of customers in the system. The associated probability distribution for the number of customers in an $M/M/1$ system in steady-state is $p_n = (1 - \rho)\rho^n$.

Now that the steady-state probability distribution has been established, the measures of effectiveness for the $M/M/1$ queuing system may be calculated. The expected value of the number of customers in the system in steady-state, W, is

$$W \equiv E\{N\} = \frac{\lambda}{\mu - \lambda} \tag{2.5}$$

The expected value of the number of customers in the queue in steady-state, L_q, is

$$W_q \equiv E\{N_q\} = \frac{\lambda^2}{\mu(\mu - \lambda)}. \tag{2.6}$$

Finally, the expected steady-state wait time for a customer in the system, W, and the expected steady-state wait time for a customer in the queue, W_q can be found using the Little formulae. Respectively, they are:

$$\tau = \frac{L}{\lambda} = \frac{1}{\mu - \lambda} \tag{2.7}$$

and

$$\tau_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu - \lambda)} \tag{2.8}$$

2.2 Hyperbolic PDEs [3]

The use of hyperbolic partial differential equations to model wave motion or advective transport of substances is so effective that this class of PDEs is often attributed the moniker of "wave" equation or "transport equation. Although, the phenomena of waves and advection arise from different physical principles, mathematically advection and wave motion (unidirectional, that is) are identical. This section presents

[3]Reference material for this section is drawn from LeVeque (1992).

some basic results from the theory of systems of hyperbolic PDEs that form the cornerstone of the forthcoming titular model.

2.2.1 General

The simplest form of a hyperbolic system of PDEs in one spatial dimension with explicit dependence on time is linear, first-order homogeneous with constant coefficients:

$$u_t(x,t) + Au_x(x,t) = 0 \tag{2.9}$$

Here $u : \mathbf{R} \times \mathbf{R} \to \mathbf{R}^m$ is an m dimensional vector representing the unknown functions that one wishs to determine and A is an $m \times m$ matrix of constants. The matrix A must have a non-degenerate eigensystem with real eigenvalues in order for this equation to be hyperbolic.

The properties of the matrix A allow us to decompose it as $A = R\Lambda R^{-1}$ where R is the matrix of right eigenvectors and Λ the corresponding vector of eigenvalues. This equation (2.9) can be solved by solving the fully decoupled system

$$w_t + \Lambda w_x = 0 \tag{2.10}$$

where $w = R^{-1}u$. The *pth* equation of (2.10) is the scalar advection equation

$$w_t^p + \lambda^p w_x^p = 0^4 \tag{2.11}$$

which has solutions of the form

$$w^p(x,t) = \bar{w}^p(x - \lambda^p t) \tag{2.12}$$

[4]Here p is not a power but an index denoting a single equation from the system.

which are the *pth* components of $\bar{w}(x) \equiv R^{-1}\bar{u}(x)$ for some given initial data $u(x,0) = \bar{u}(x)$.

The solution

$$u(x,t) = \sum_{p=1}^{m} w^p(x,t) r^p \tag{2.13}$$

illustrates why (2.9) is called the transport equation. Hyperbolicity guarantees that the form of $u(x,t)$ at the point x at time t can be seen as a linear combination of right eigenvectors $r^1, \ldots, r^p$ and thus is the superposition of waves propagating with velocities λ^p. The strength of each wave is given by the coefficients $w^p(x,t)$ and these functions are called the *characteristic variables*. As time evolves, the eigencoefficient $\bar{w}^p(x) \equiv w^p(x,0)$ is advected with constant velocity λ^p along the curve $X(t) = x_0 + \lambda^p t$. The curves X are called the *characteristics of the pth family* and in the case of constant A are straight lines.

More generally, consider the system

$$u_t + A(x)u_x = 0 \tag{2.14}$$

where the matrix $A(x)$ now depends explicitly on the position x. If $A(x)$ is diagonalizable with real eigenvalues for each x in a domain, then the system (2.14) is hyperbolic in that domain. The solution to this system in the hyperbolic domain can again be written as a linear combination of right eigenvectors, but in this case the characteristics are not straight lines and rather they are solutions to a system of ordinary differential equations (ODEs) derived from the decoupled system for $w(x,t)$.

2.2.2 Conservation Laws

An important subclass of hyperbolic PDEs are those called *hyperbolic conservation laws*. These are homogenous hyperbolic PDEs that model that conservation of one, or many, quantities of interest.

One important quantity that is conserved in many physical problems- and in this research also- is that of mass. Suppose there is a distribution of mass in one spatial dimension x that is subject to an advection motion traveling in the positive x direction which has a velocity profile $\nu(x,t)$ at the point x at time t. Define a density $\rho(x,t)$ at point x and time t in a given interval $[x_1, x_2]$ in the following manner

$$mass\ in\ [x_1, x_2]\ at\ time\ t = \int_{x_1}^{x_2} \rho(x,t)dx \tag{2.15}$$

Then, assuming that there are no mass sources or sinks, the mass in this interval may only change via flow through the endpoints x_1 or x_2. The $flux$, $F(x,t)$, is the rate of flow of mass at the point x at time t. That is

$$F(x,t) = \rho(x,t)\nu(x,t) \tag{2.16}$$

Therefore, over the interval $[x_1, x_2]$, the rate of change of mass is given by

$$\frac{d}{dt}\int_{x_1}^{x_2} \rho(x,t)dx = F(x_1,t) - F(x_2,t) = \rho(x_1,t)\nu(x_1,t) - \rho(x_2,t)\nu(x_2,t) \tag{2.17}$$

Equation (2.17) is known as the *integral form* of the mass conservation law.

Assuming that both ρ and ν are differentiable functions of x and t, one can derive the *differential form* by observing that for arbitrary x_1, x_2, t_1, and t_2 the equation

$$\int_{t_1}^{t_2}\int_{x_1}^{x_2} \left\{ \frac{\partial}{\partial t}\rho(x,t) + \frac{\partial}{\partial x}(\rho(x,t)\nu(x,t)) \right\} dxdt = 0 \tag{2.18}$$

must hold true. Hence,

$$\rho_t + (\rho\nu)_x = 0 \tag{2.19}$$

as the resulting differential form of the conservation law for mass.

More generally, in one spatial dimension, x, and time dimension, t, the simplest hyperbolic system of conservation laws takes the form

$$u_t(x,t) + F(u(x,t))_x = 0. \tag{2.20}$$

Again $u : \mathbf{R} \times \mathbf{R} \to \mathbf{R}^m$ is an m dimensional vector representing the unknown functions of conserved quantities that one wishes to determine and the vector-valued function $F : \mathbf{R}^m \to \mathbf{R}^m$ is the $flux$ of these quantities. Rewriting equation (2.20) in *quasilinear* form gives

$$u_t + F'(u)u_x = 0. \tag{2.21}$$

Equation (2.21) will be hyperbolic if the Jacobian $F'(u)$ has real eigenvalues corresponding to a linearly independent set of eigenvectors. This is not unlike the constant or variable coefficient case.

2.2.3 Nonlinearity and Burger's Equation

While the differentiability of $u(x,t)$ was assumed in the derivation of the differential form of hyperbolic conservation laws, it turns out that spatial smoothness is not a necessary requirement to construct a "solution" to the PDE. As illustrated, the solution $u(x,t)$ along a characteristic depends on one value $\bar{u}(x)$. For linear hyperbolic equations, singularities in $\bar{u}(x)$ are maintained with the same order by ($u(x,t)$ and propagate along the characteristics. Nondifferentiability in $\bar{u}$ at some point x preludes $u(x,t)$ from being a classical solution the the differential equation everywhere. Yet,

$u(x,t)$ does still solve the integral form of the conservation law. Functions $u(x,t)$ that satisfy the integral form, but not necessarily the differential form of the conservation law are know as *weak solutions* [5] .

The integral form of the conservation law (2.17) has a more fundamental physical basis than the differential form (which is derived from it) and nonsmoothness of initial data does not invalidate the efficacy of the model. Yet, solving the integral form is more difficult than solving its corresponding differential form. One way in which the generalized solution can be constructed for nonsmooth initial data $\bar{u}(x)$ is by approximating such data with a sequence of smooth functions $\bar{u}^\epsilon(x)$ where

$$\|\bar{u} - \bar{u}^\epsilon\|_1 < \epsilon \tag{2.22}$$

as $\epsilon \to 0$. Here $\|\cdot\|_1$ is the L_1 norm.

The linear PDE along with smooth initial data $\bar{u}^\epsilon$ will have a smooth classical solution $u^\epsilon(x,t)$ for all $t \geq 0$. One can then take as the generalized solution the limit

$$u(x,t) = \lim_{\epsilon \to 0} \bar{u}^\epsilon(x,t) \tag{2.23}$$

This procedure for smoothing works well for linear hyperbolic PDEs, but smoothing the initial data will not work for nonlinear problems. One can, though, modify (2.9) by adding a small diffusive term and observe that the original conservation law can be regarded as an approximation to the advection-diffusion equation

$$u_t + Au_x = \epsilon u_{xx} \tag{2.24}$$

for ϵ very small. The solution to (2.24), $u^\epsilon(x,t)$, with initial data $\bar{u}(x)$ is the an element of $C^\infty((-\infty,\infty) \times (0,\infty))$ regardless of the smoothness of $\bar{u}(x)$ since the PDE

[5] In some literature these are also known as *generalized* solutions

is parabolic. The generalized solution is then obtained by taking the limit of $u^\epsilon(x,t)$ as $\epsilon \to 0$ as before. This type of approach can work for nonlinear problems.

Consider the nonlinear scalar equation

$$u_t + F(u)_x = 0 \tag{2.25}$$

where the flux function is given by

$$F(u) = \frac{1}{2}u^2 \tag{2.26}$$

This equation is known as inviscid Burger's equation. It is about the simplest model including the nonlinear, viscous effects of fluid dynamics. It also illustrates an issue that is not found in linear equations: Nonlinear equations can generate discontinuous solutions even with smooth initial data.

2.2.4 Riemann Problem

The Riemann problem is a conservation law combined with piecewise constant initial data having a single discontinuity. Taking Burger's equation, for example,

$$u_t + uu_x = 0 \tag{2.27}$$

together with initial data

$$u(x,0) = \bar{u}(x) = \begin{cases} u_l & \text{if } x < 0, \\ u_r & \text{if } x > 0 \end{cases} \tag{2.28}$$

the general solution has a form that depends on the relationship between u_l and u_r.

Case I. $\mathbf{u_l < u_r}$

In this case there is a weak solution called a *shock wave* which has the form

$$u(x,t) = \begin{cases} u_l & \text{if } x < st, \\ u_r & \text{if } x > st \end{cases} \tag{2.29}$$

where

$$s = \frac{u_l + u_r}{2} \tag{2.30}$$

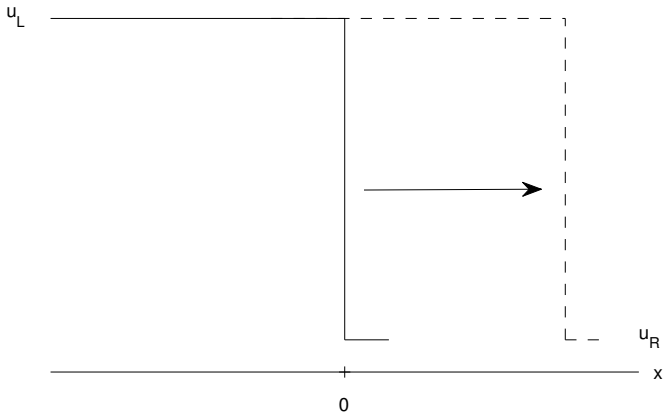
is the *shock speed*. The shock speed is the speed at which the discontinuity travels and the characteristics in each region where u is constant propagate into the shock. Figure 2.2 is an example of such a shock solution.

Case II. $\mathbf{u_l > u_r}$

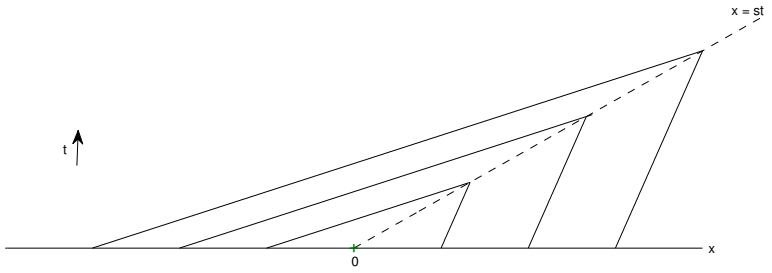
There are infinitely many weak solutions to (2.27) in this case and one of them happens to be (2.29), (2.30). However, this solution is not stable to perturbation in this case; any added viscosity or slight smearing of the initial data will completely change the solution. A stable weak solution to (2.27) is the *rarefaction wave*

$$u(x,t) = \begin{cases} u_l & \text{if } x < u_l t \\ x/t & \text{if } u_l t < x < u_r t \\ u_r & \text{if } x > u_r t \end{cases} \tag{2.31}$$

This solution also happens to be the vanishing viscosity generalized solution. Figure 2.3 illustrates an example rarefaction wave solution.

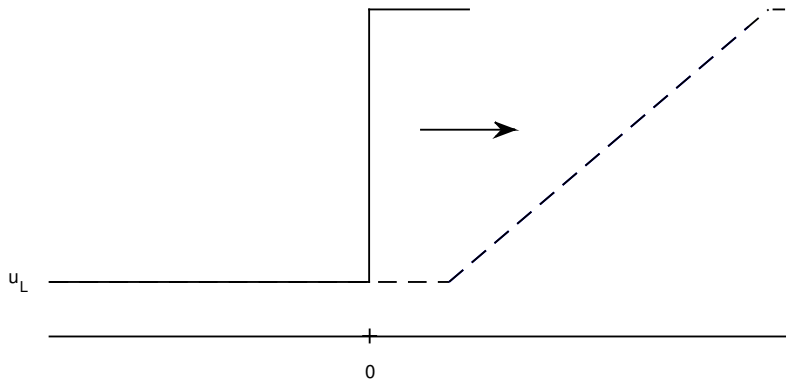


(a) Progression of Shock.

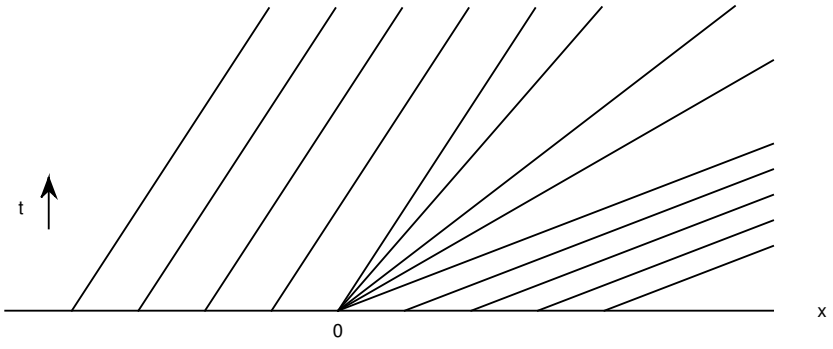


(b) Characteristics in Phase Space

Figure 2.2: Shock Wave for Solution 2.29



(a) Progression of Rarefaction Wave.



(b) Characteristics in Phase Space

Figure 2.3: Shock Wave for Solution 2.31

More generally, for arbitrary flux function $F(u)$ the relationship between the two constant states u_l and u_r and the shock speed s is encapsulated by the *Rankine-Hugoniot (R-H) jump condition*:

$$F(u_l) - F(u_r) = s(u_l - u_r) \tag{2.32}$$

For scalar equations, solving for s is quite easy, yet for systems of equations, $F(u_l) - F(u_r)$ and $u_l - u_r$ are vectors with scalar s. It is not always possible to retrieve s from the R-H condition (2.32) unless we restrict the types of jumps allowed at the discontinuity, namely to those for which the above vectors are linearly independent.

The validity of the R-H conditions (2.32) is not restricted to piecewise constant initial data. These results hold when considering a propagating shock with initial data the is smooth in the regions to the left and right of the discontinuity. In this case, the immediate values to the left and right of the discontinuity are denoted as u_l and u_r, respectively.

Chapter 3

SIMULATION

Effective simulation is an immediate issue when investigating stochastic phenomena. In order to be effective, the simulation must both accurately recreate the probabilistic mechanisms of the model and generate the record of the quantities of interest as they evolve over time. However, even a relatively simple probabilistic model can suffer from a complex logical structure of its elements that precludes an obvious strategy for tracking the model's evolution and, consequently, the values of the quantities desired. An approach that has become increasingly popular is built around the idea of "discrete events". This type of approach is called a *discrete event simulation (DES)* approach. This chapter discusses the basics of a specific DES program utilized in this research known as χ and how the research incorporates this approach as the baseline for testing the efficacy of the second order transport model. For a more comprehensive discussion of discrete event simulation the interested reader is referred to Ross (2013) or Kelton and Law (2000).

In Chapter 2 the $M/M/1$ queueing system was introduced. For constant arrival and service (machine) rates, this system is completely understood and one can determine many of its salient features, especially in the steady-state. Yet, constant rate behavior is neither of much interest from modeling perspective nor is it very useful in a large variety of real-world situations. What is more pertinent is the behavior of such a queueing system as the arrival and/or service rate vary in time. Since this behavior is less well understood reliance on simulation as a representation of the true behavior of the system is taken.

This research focuses on a selection of three distinct, non-constant arrival (influx) patterns: (1) Exponential Relaxation; (2) Finite jump (the Step); (3) Periodic (the Cycle). In this chapter, the real-world motivation for these influx patterns will be discussed and the important features of the resulting departure (outflux) patterns investigated. Additionally, an overview of the techniques employed to construct the simulations and generate the data will be provided.

3.1 Discrete Event Simulation

The $M/M/1$ queueing system is an example of a birth-death process. The defining characteristic of such a process is that the state of the system changes only through a birth (arrival) or a death (departure). Consequently, one can obtain a complete picture of the system as it evolves over continuous time by only tracking those distinct state changing events and the times at which these events occur. These discrete events are recorded in an *event list* and one can simulate an $M/M/1$ queueing system by generating such a list. Naturally, this type of simulation is called a discrete event simulation (DES) and it is the type of simulation technique employed in this research.

The DES approach has at it's core two key elements: (1) variables; (2) events. As to the variables, there are three that are generally encountered, namely, a time variable, a counter variable and a system state variable. The *time* variable is used to track the amount of time (simulated) that has elapsed since the start of the simulation. The *counter* variable is used to label the number of times that a certain event has occurred or as an index in a sequence of successive events.. The *system state* variable records the state of the system at specified times.

For most DES the *events* are changes in the state and counter variables. When an event occurs, the current time is recorded and the variables updated. One also may record at this time additional data such as output, input, etc. Since these events occur at punctuated times, rather than over intervals they are discrete hence the simulation's moniker. Moreover, since the systems do not change until an event, record keeping costs can be minimized without a loss of information. Virtually any birth-death process is amenable to accurate simulation via discrete events and as a birth-death process, the $M/M/1$ is no exception.

While there are some significant differences among the various software programs used in DES, at their core, they are functionally equivalent. Recall Figure 2.1. Events are generated by two processes separated by a queue (buffer) thats stores the arriving lots until they can enter the machine:

A: arrival process: Let t_a be the time of the current arrival. Then at $t = t_a$ two things happen: 1) a lot is sent to the queue; 2) a new lot is created and a new inter arrival time τ_A is pulled out of the exponential distribution with parameter $1/\lambda$. The next arrival event is then calculated as $t_a = t_a + \tau_A$ and put into the event list.

M: service (machine) process: Let t_e be the time that the current lot in service exits the server. Then at time $t = t_e$ the following things happen: 1) The part currently in the machine exits the machine (either to the exit process or to another machine or queue); 2) The machine becomes available again; 3a) If the queue is empty, the machine idles until the queue is non-empty again. Then continue with 3b); 3b) If the queue is non-empty, the machine pulls the next part from the queue as determined by the queue priority (if the queue priority is FIFO, then this would be the oldest part); 4) The machine pulls a service time τ_M out of the exponential distribution with

parameter $1/\mu$. The next exit event is then calculated as $t_e = t_e + \tau_B$ and is put into the event list.

Now at the end of every arrival or machine process an event occurs and the event list is updated and reordered chronologically starting with the time of the most recent event. The process that generated the event is then restarted. Both the arrival process and machine process are run simultaneously if they are able. The arrival and machine processes are easily the most important aspects of any DES as the buffer - in it's simplest form - is nothing more than a storage array.

3.1.1 Generating a Poisson Process

Because of their paramount importance in the dynamics of the $M/M/1$ queueing system we must take care to simulate the arrival and machine processes in a manner that is consistent with the influx and service profiles that we wish to represent. Our approach to the simulation will be very dependent on the time varying nature of these rate profiles with the arrival rate λ generally varying in time and the service rate μ always constant.

Consider the arrival process with parameter $\lambda > 0$. For constant λ this process is said to be a *homogeneous* or *stationary* Poisson process. Formally, let $N(t)$ $(t \geq 0)$ be a stochastic process that tracks the number of events that have occurred up to time t. Such a process is called a counting process. Then the homogeneous Poisson process is defined as:

Definition 3.1 *A homogeneous Poisson process N(t) is a counting process with the following additional properties:*

Step	Subroutine
1	$t = 0$, $i = 0$.
2	Generate random number $t_a \sim \text{Exp}(1/\lambda)$.
3	$t = t + t_a$. If $t > T$, break.
4	$i = i + 1$, $E(i) = t$.
5	Return to Step 2.

Table 3.1: Algorithm for Creating the Event List E for a Homogeneous Poisson Process with Rate λ.

- $N(0) = 0$
- *Stationary increments - the number of events in a given interval depends only on the length of the interval*
- *Independent increments - the number of events in disjoint intervals is independent*
- $N(t) \sim Poisson(\lambda t)$

Simulating such a process is quite easy and Table 3.1 provides a general algorithm for an event list E using DES for the first T time units. In the algorithm, i is the counter variable denoting the number of events occurring by time t and $E(i)$ is the ith event time.

Relaxing the Poisson process assumption of stationary increments by letting the arrival rate vary with time yields a *nonhomogeneous* Poisson process:

Definition 3.2 *A nonhomogeneous Poisson process N(t) is a counting process with the following additional properties:*

- $N(0) = 0$

- *Independent increments - the number of events in disjoint intervals is independent*

- $N(t) \sim Poisson(\int_0^t \lambda(s)ds)$

In simulating a time varying influx $\lambda(t), 0 \leq t \leq T$, it might seem intuitive to replace λ with the desired function $\lambda(t)$ in the above algorithm Table 3.1. That is, given an arrival at time t_i generate the next arrival by sampling from the distribution $\text{Exp}(\lambda(t_i))$ instead of $\text{Exp}(\lambda)$ like in the constant case. However, this is not a valid approach. [1] A standard approach is that given by Lewis and Shedler (1978) called the *thinning* or *random sampling* method. An important property of Poisson queues should be remarked on first, though.

Poisson-expoenential arrival processes are often referred to as completely random arrivals in queueing literature. This colloquialism suggests that a uniform distribution is somehow involved which appears to be inconsistent with the Poisson-arrival-rate-expoenetial-interarrival-time pattern. Yet it is true; a uniform distribution is involved in this stochastic process. It is the times at which the arrivals occur that is uniformly distributed, not the interarrival times. Proof of this can be seen by examining the conditional probability density's differential element and applying the definition of conditional probability.

Recall that the order statistics of k uniform random variables on $[0, T]$ has a joint density of $k!/T^k$. Let $\tau_1 < \tau_2 < \cdots < \tau_k$ be the corresponding times of k arrivals that have occurred in an interval [0, T]. Then

[1] This will be illustrated in the comparison of the exponential and step scenarios to come.

$$
\begin{aligned}
f_\tau(t|k)dt &\equiv f(t_1, t_2, \ldots, t_k | k \text{ arrivals in } [0,T])dt_1 dt_2 \cdots dt_k \\
&\approx \Pr\{t_1 \le \tau_1 \le t_1 + dt_1, \ldots, t_k \le \tau_k \le t_k + dt_k | k \text{ arrivals in } [0,T]\} \\
&= \frac{\lambda e^{-\lambda dt_1} dt_1 \lambda e^{-\lambda dt_2} dt_2 \ldots \lambda e^{-\lambda dt_k} dt_k e^{-\lambda T - dt_1 - dt_1 - \cdots - dt_k}}{(\lambda T)^k e^{-\lambda T}/k!} \\
&= \frac{k!}{T^k} dt_1 dt_2 \cdots dt_k
\end{aligned}
$$

Hence,

$$
f_\tau(t_1, t_2, \ldots, t_k | k \text{ arrivals in } [0,T]) = \frac{k!}{T^k}.
$$

The above uniform property of the Poisson process is an important and well known one in queueing theory as it gives rise to the heavily utilized $PASTA$ ("Poisson arrivals see time averages") property for queues. If a stochastic process $X(t)$ is a queueing system then the expected value of any parameter of the system [2] as seen by a Poisson arrival will be equivalent to the long run average value of that parameter.

The thinning method takes advantage of the fact that arrivals governed by a homogeneous poisson process have arrival times that are uniformly distributed. This is done by randomly selecting the arrival event times of a homogeneous Poisson process having rate λ^* with probability $\lambda(t)/\lambda^*$. In other words, an arrival at time t is selected with probability $\lambda(t)/\lambda^*$ and unselected arrivals are discarded. The choice for λ^* depends on the simulation of interest, but since this research restricts $\lambda(t)$ to be bounded, a simple and efficient choice is to let $\lambda^* = \max(\lambda(t)), 0 \le t \le T$.

[2] Recall that these are often the measures of effectiveness of the queues as discussed in Chapter 2.

Step	Subroutine
1	$t = 0$, $i = 0$.
2	Generate random number $t_a \sim \text{Exp}(1/\lambda)$.
3	$t = t + t_a$. If $t > T$, break.
4	Generate random number $x \sim \text{Uniform}(0,1)$
5	If $x \leq \lambda(t)/\lambda$, set $i = i + 1$, $E(i) = t$.
6	Return to Step 2.

Table 3.2: Algorithm for Creating the Event List E for a Nonhomogeneous Poisson Process with Rate $\lambda(t)$.

In implementation, a stationary Poisson process is generated with constant rate $\lambda = \lambda^* = \max\{\lambda(t)\}$ and arrival times $E(i) = t_i$, then "thin out" the t_i's by only accepting into the final event list t_i with probability $\lambda(t_i)/\lambda$. The result is that t_i's have a higher likelihood of being accepted as an arrival event if $\lambda(t_i)$ is high and are rejected with a higher probability if $\lambda(t_i)$ is low. This reflects the desired property that arrivals will occur with greater frequency in intervals for which $\lambda(t)$ is high as opposed to when it is low. A simple and convenient recursive algorithm is provided in Table 3.2(we retain the same notation as is Table 3.1). [3]

As mentioned, the thinning method is quite simple and very efficient for the arrival rates used in this research. Figure 3.1 shows a sample run using this method for the piecewise constant $\lambda(t)$ (green) given by

[3]For brevity, we assume the the arrival t_{i-1} has been validly generated and focus just on the generation of the next arrival t_i.

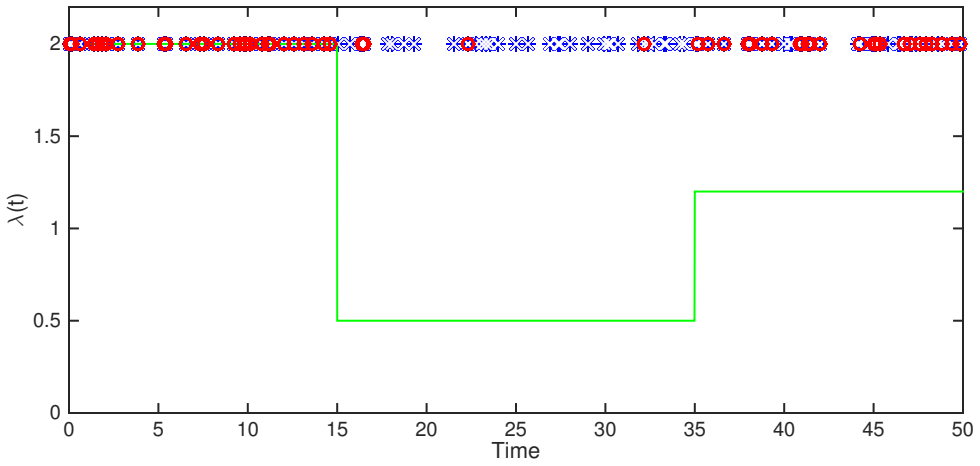


Figure 3.1: Example of Thinning Method for Single Run with $\lambda(t)$ Given by (3.1).

$$\lambda(t) = \begin{cases} 2 & 0 \leq t < 15, \\ 0.5 & 15 \leq t < 35, \\ 1.2 & 35 \leq t \end{cases} \tag{3.1}$$

The blue stars represent arrivals corresponding to the homogeneous Poisson process with rate parameter $\lambda^* = 2$. Those stars that are encircled (red) are the arrivals that have been accepted. As can seen from this diagram all arrivals occurring before $t = 15$ are accepted with probability 1. For $t \in [15, 35)$ the acceptance probability drops to 0.25 $(\lambda(t)/\lambda^*)$ and the number of acceptances is greatly reduced. Finally, for $t \geq 35$ the acceptance probability increases to 0.6 and more arrivals are accepted then in the previous regime, albeit not as much as in the first interval, as desired.

3.2 Experimental Scenarios and χ

This chapter closes out with a description of the three experimental scenarios against which the second order model was tested. The DES coding program will

be discussed in brief detail along with how it was used to generate the production flows of the $M/M/1$ queue. Finally, the notable features of the simulations for these scenarios will be illustrated and will be revisited in later chapters.

Each of these scenarios has a foundation in real world factory dynamics. Obviously, the $M/M/1$ queue is a highly simplified stand-in for a factory dynamic, however many behaviors and processes of factories when examined in aggregate have prominent features in common with queueing systems. Arrival rates and internal service rates of the factory are stochastic. Machine failures, labor issues and production resource disruptions etc. are common examples of random influences affecting the flow of production through a factory. No specificity to any particular influences of interest will be made, however. Rather, these influences and others are incorporated in the stochastic processes used for arrival rates and service rates. Additionally, production resources often are not immediately passed into service upon their arrival. Rather, they are taken up only when service is available. Local inventory systems are implemented in most factories to manage the practical logistics due to the time difference between the arrival events and their ability to enter service. Naturally, these inventory systems are the factory queues and this is what queueing system models built to represent.

The choice for Poisson as the arrival and service process is not entirely for mathematical ease, although it is surely an attractive feature given the breadth of theory pertaining to the $M/M/1$ queue and Poisson processes in general. It is a reasonable assumption that for generalized factories of sufficient size, the stochastic influences are not only independent of each other, but that they are also memoryless. A machine going down does not, in general, affect the rate of another machine performing the

same task on a different production line. Neither does a production resource shortage become more or less probable as the time since that last shortage occurred increases. Of course, this assumption is not always the case, but it is ubiquitous enough that using Poisson processes as an aggregate approximation of factory arrival and service rates is not wholly academic.

In this paper, the simulated $M/M/1$ queuing system generated by the computer program χ (Chi) is considered as the baseline against which the hyperbolic continuum model is compared. The creation of one event list is called a *run* and the total number of runs for a given input pattern is called a *simulation*. In order to derive useable statistics, a histogram is built with the runs in the simulation. The simulated-time interval [0, T] for the DES is partitioned into equispaced intervals of width $\Delta t = 0.1$. Over each interval $[t_i, t_{i+1})$ the number of specific events that occurred in this interval is counted. [4] This is then divided by the total run count of that particular simulation. This histogram is then used as the representative profile for the flow though the queueing system. In each of the following scenarios, the statistics are gathered over a sample size of 40,000 runs per simulation.

3.2.1 Scenario 1: Exponential Relaxation

The first scenario considered is the exponential relaxation input pattern. Suppose that the initial lot arrival rate of λ_1 constant. At a prescribed time T_0 the lot arrival rate is changed to a new constant rate λ_2 but this rate only affects those lots subsequent to the first arrival post T_0. In other words, if t_i is the last arrival event before

[4] By specific, we mean that if we are generating the outflux pattern, for example, we count only the corresponding exit events of the simulation.

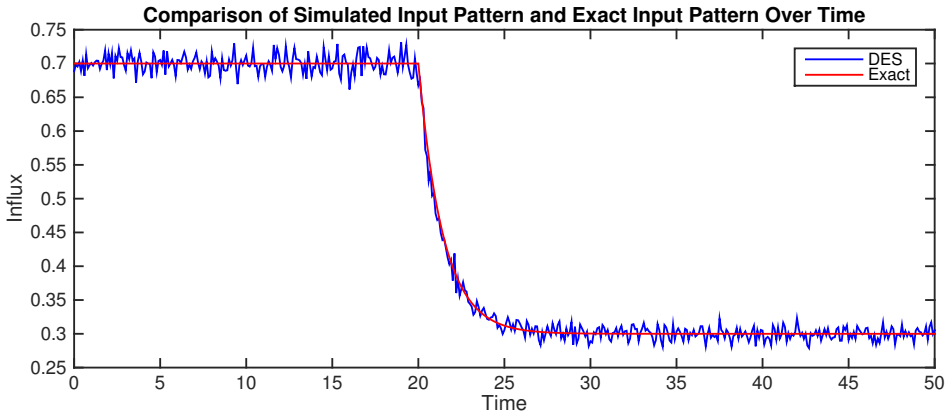
time T_0 then arrival t_{i+1} occurs after T_0 with an interarrival time corresponding to rate λ_1 but all arrival events after t_{i+1} have an interarrival time depending on λ_2. This setup gives an arrival pattern of the following form:

$$\lambda(t) = \begin{cases} \lambda_1 & \text{if } t \leq T_0 \\ (\lambda_1 - \lambda_2)e^{-\lambda_1(t-T_0)} + \lambda_2 & \text{if } T_0 < t \end{cases} \tag{3.2}$$

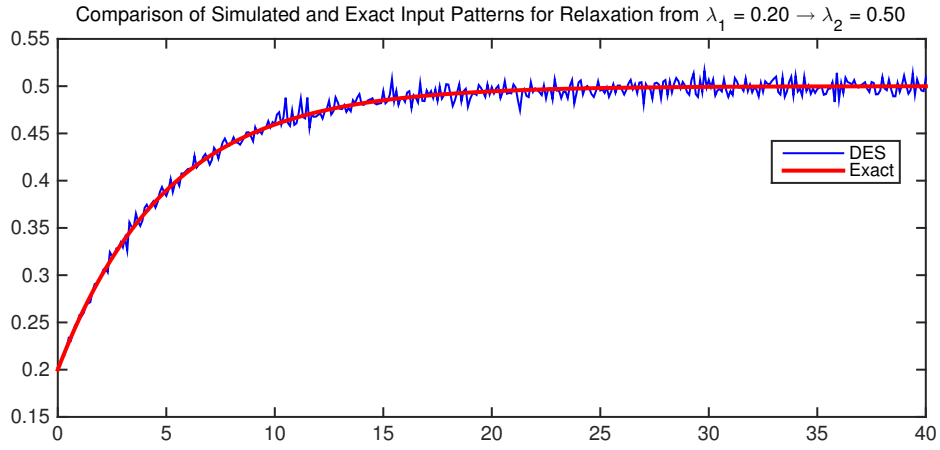
since it is necessary that $\lim_{t\to\infty} \lambda(t) = \lambda_2$ and $\lambda(T_0) = \lambda_1$. Figure 3.2 provides an illustration of $\lambda(t)$ for a typical increasing and decreasing transition alongside the relevant DES output data.

As an aside, it should be noted that one is not required to use a nonhomogeneous Poisson process simulation approach in this case even though $\lambda(t)$ is non constant. This is because $\lambda(t)$ has been constructed to be piecewise constant with a single jump discontinuity. One could, therefore, simulate the arrival process by first simulating a homogenous Poisson process with rate λ_1 for arrivals up to and including arrival t_{i+1}, simulating arrivals after t_{i+1} with rate λ_2, and then creating the arrival event list from a concatenation of these individual lists.

For a factory the exponential relaxation input pattern could represent a decision by management to reduce the influx to the new level λ_2 at time T_0 but is restricted in practice from doing so immediately. This delay could be due to supply contracts that must be fulfilled at the original level or that arrivals are the output of other production facilities which are, for some reason, unable to alter their production schedule until their current WIP is completed. This manifests itself in a slow decay/growth to the new rate level λ_2 as individual production resource streams switch over to this new level in time.



(a) Decreasing Transition ($\lambda_1 = 0.70$, $\lambda_2 = 0.30$, $T_0 = 20$)



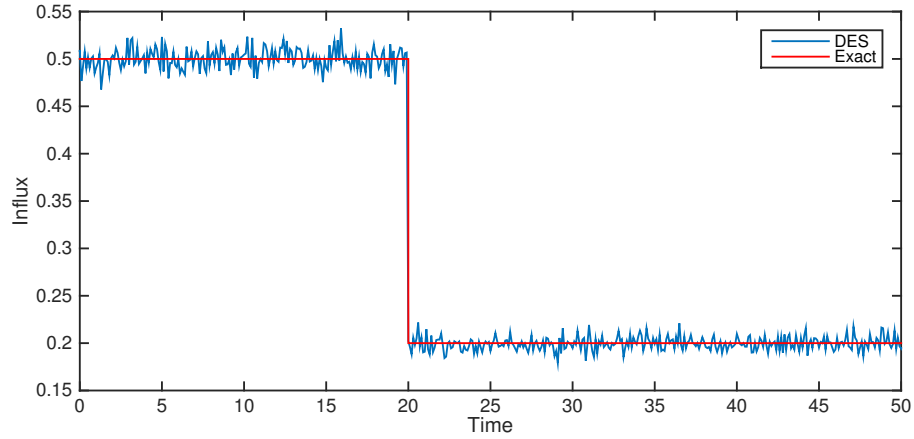
(b) Increasing Transition ($\lambda_1 = 0.20$, $\lambda_2 = 0.50$, $T_0 = 0$)

Figure 3.2: DES and Exact Input Patterns for Exponential Relaxation Scenario.

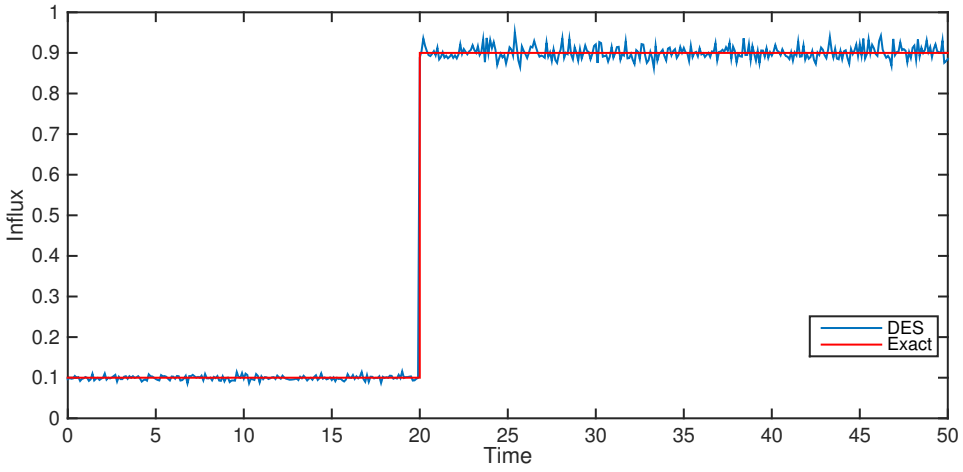
3.2.2 Scenario 2: The Step

The step scenario is virtually identical in setup to the exponential relaxation case with a significant exception: At T_0 the rate $\lambda(t)$ changes from λ_1 to λ_2 without delaying until arrival t_{i+1}. This yields a true step-like behavior in $\lambda(t)$ as seen in Figure 3.3 for representative increasing/decreasing transitions.

Unlike in the exponential case, in order to produce a valid DES for this scenario one is required to use a simulation approach such as thinning that is tailored to non-



(a) Decreasing Transition ($\lambda_1 = 0.50$, $\lambda_2 = 0.20$, $T_0 = 20$)



(b) Increasing Transition ($\lambda_1 = 0.10$, $\lambda_2 = 0.90$, $T_0 = 20$)

Figure 3.3: DES and Exact Input Patterns for Stepwise Transition Scenario.

homogenous Poisson processes even though these scenarios are very similar. Point in fact, if an attempt to simulate this type of nonhomogeneous Poisson arrival process was made in the naive manner mentioned earlier in the chapter by simply replacing λ with $\lambda(t)$ the result would be the previous exponential pattern. This is because the arrival t_{i+1} would have an arrival rate dependent on $\lambda(t_i) = \lambda_1$ while subsequent arrivals are dependent on the rate $\lambda(t_{i+1}) = \lambda(t_{i+2}) = \cdots = \lambda_2$.

In a factory, an example of how a step type input pattern could result would be in the case of a partial cutoff of the production resource. This should be considered separate from the supply disruptions that give rise to the stochasticity of the arrival process and are aggregated into the arrival flow.

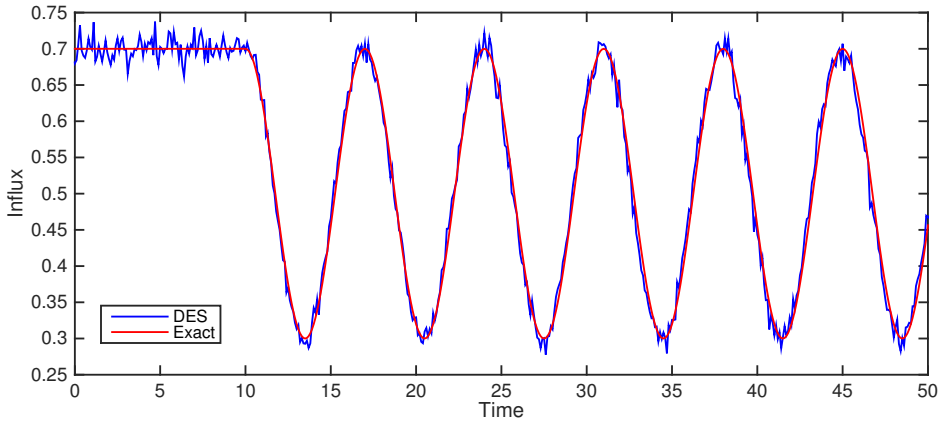
3.2.3 Scenario 3: The Cycle

The final scenario that is considered is also one with a nonhomogeneous arrival rate. It is the cyclic input pattern. Starting from an initial constant rate λ_1 at some prescribed time T_0 the input pattern $\lambda(t)$ steadily oscillates with a given amplitude between the initial rate λ_1 and a minimum rate λ_2 . Precisely, the arrival rate is given by

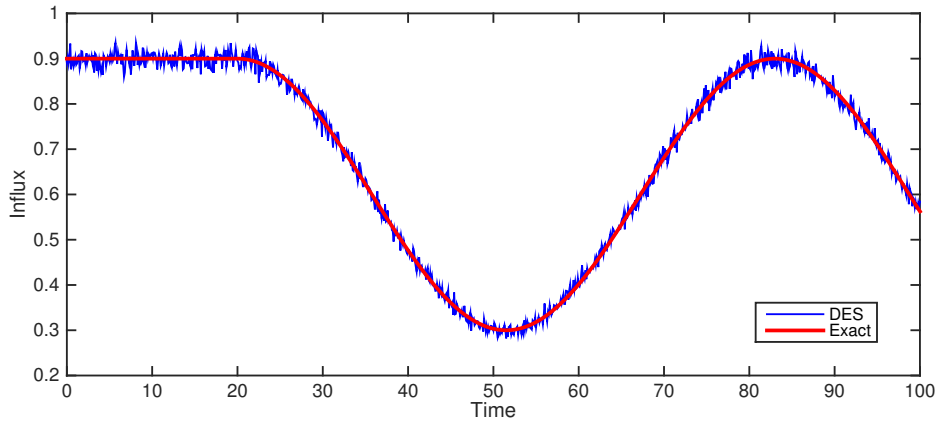
$$\lambda(t) = \begin{cases} \lambda_1 & \text{if } t \leq T_0 \\ \frac{\lambda_1 - \lambda_2}{2}\cos(\frac{2\pi t}{C\tau}) + \frac{\lambda_2 + \lambda_1}{2} & \text{if } T_0 < t \end{cases} \tag{3.3}$$

where $\tau = 1/(\mu - \lambda_1)$ is the steady-state lot cycletime corresponding to the initial arrival rate λ_1 and service rate μ. The additional parameter C is used to generate DES inputs for various periods yet with the same amplitude.

Cyclic behavior is widespread in industry. In many industries, the supply of pro-



(a) $C = 3$, Range [0.3 0.7]



(b) $C = 7$, Range [0.3, 0.9]

Figure 3.4: DES and Exact Input Patterns for Cyclic Scenario.

duction resources varies periodically over time. For simplicity, elementary cosine and sine functions were chosen for the cyclic behavior. In addition, from a mathematical perspective, more complicated input patterns may be approximated with linear combinations of these functions and an analysis of this hyperbolic model with these input patterns is a necessary first step. Figure 3.4 shows a few of the experimental input patterns.

Chapter 4

THE CONTINUUM MODEL

There has long been a tradition of aggregating the stochastic flow of products through a factory in two ways:

- averaging over time or over ensembles to convert the stochastic process into a deterministic production process and
- converting work in progress (WIP) from integer variables, labeling individual parts, into a real variable describing a product continuum like a fluid.

Together these two aggregation steps convert queuing network models into ordinary differential equation (ODE) models known as *fluid models*. Queuing networks are analyzed using network equations linking the random variables describing the state of the network. *Fluid equation models* are the deterministic equations replacing the random variables with their means. Fluid models conceptually arise from treating the jobs in a queuing network as a continuous fluid that flows, via inflow and outflow rates, through a finite number of reservoirs (the queues),

$$\frac{dq_i}{dt} = \begin{cases} \mu_{i-1} - \mu_i & \text{if } q_i \neq 0, \\ 0 & \text{if } q_i = 0 \end{cases} \tag{4.1}$$

for $i = 1, \ldots, N$ and $\mu_0 = \lambda$, where q_i is the WIP waiting for step i, μ_i is the processing rate of step i, and λ is the start or arrival rate into the production process.

The resulting models are hybrid dynamical systems: sets of ODEs for the time evolution of the queue lengths as a function of time. The appeal of fluid models is that they are deterministic dynamical systems that are well understood even though some important issues related to the stability of queuing networks and the stability of the associated fluid models remain unresolved for multi class queuing systems (Bramson (2008), Dai (1995), Dai *et al.* (2004)).

Fluid models do not really behave like a fluid, because they still treat every production process separately and hence model the production flow through discrete steps. For long production lines with many steps, it makes sense to treat the production steps as a continuum variable and in that way obtain a genuine fluid dynamical description that treats the factory as a pipe and the parts flowing through the factory as a fluid. In contrast to a real fluid, the spatial variable does not describe physical space, rather it denotes the degree of completion of the part, that is, how far along the part is in the system. Calling $x \in [0, 1]$ the degree of completion (where $x = 0$ and $x = 1$ denote recent arrivals and departures, respectively), $\rho(x, t) \geq 0$ describes the density of parts at stage x at time t, and $W(t) = \int_0^1 \rho(x, t)dx$ is the total WIP in the factory. If the fluid moves with a velocity $\nu(x, t)$, then the flux is described as $F(x, t) = \rho(x, t)\nu(x, t)$. Assuming that the defective products are sorted out after the factory production process, there are no sources or sinks in a factory, the WIP satisfies the mass conservation law:

$$\frac{dW}{dt} = \lambda - \mu, \tag{4.2}$$

where μ is the overall mean production rate of the factory. By a standard argument of transport equations LeVeque (1992) this integral conservation law is equivalent to

a differential conservation law of the form

$$\frac{\partial \rho}{\partial t} + \frac{\partial F}{\partial x} = 0. \tag{4.3}$$

Because $\nu(x,t) \geq 0$, the fluid moves from left to right. Hence, the boundary condition is set as the influx $F(0,t) = \lambda(t)$; i.e. the local flux at stage zero is the arrival rate of the parts into the factory. Together with an initial WIP profile $\rho(x,0) = \rho_0(x)$ and, this sets up a well-defined transport equation (hyperbolic) problem.

Using transport equations is an immediate improvement over the use of a clearing function. A delay between the influx and the outflux in the factory is automatically built into the model. The crucial modeling part affecting the lead time is the associated flux model:

- A constant velocity like in the previous example corresponds to a constant delay between the start and finish of an time, i.e. a constant lead time.

- A local flux at position x and time t depends only on the density around that position x.

  This is typically used in traffic flows and in its simplest form look like

$$F(x,t) = \nu(\rho)\rho = \nu_0(1 - \frac{\rho}{R})\rho \tag{4.4}$$

  Equation (4.4) is known as the Lighthill-Whitham model and reflects the fact that drivers slow down as the density of cars around them increases. That process continues until the density becomes critical, $\rho_c = R$, and the velocity goes to zero, indicating a traffic jam.

- A global flux is used for a model that treats the whole factory as a single queue. As a result, the velocity at position x at time t depends on the global quantity of total WIP, i.e. $\int_0^1 \rho(x,t)dx$. For instance, an $M/M/1$ queue with arrival rate λ and service rate μ can be analyzed completely in steady-state. The resulting characteristics for the relationships between cycle time τ, queue length W, and influx are

$$F(\rho(x,t)) = \frac{\mu\rho(x,t)}{1+\int_0^1 \rho(x,t)dx} = \frac{\mu\rho(x,t)}{1+W(t)}, \tag{4.5}$$

  which in steady-state reduces to a clearing function of the type

$$F(W) = \frac{\mu W}{1+W}. \tag{4.6}$$

Hence a PDE model combining a transport equation with an $M/M/1$ flux model makes the following assumptions about the production process:

- The velocity of transport, i.e. the speed at which a part moves through the production line, depends on the amount of WIP in the factory.
- The velocity at a given time is the inverse of the cycle time for an $M/M/1$ queuing system with the steady-state WIP equal to the current total load in the factory.
- The velocity is the same at every production stage, depending on the total WIP in the factory.

A slight generalization of this approach leads to a very usable model: By determining the state equation for the velocity $\nu = V(W)$, either through more elaborate queuing theory models, through measurements in the factory, or through detailed

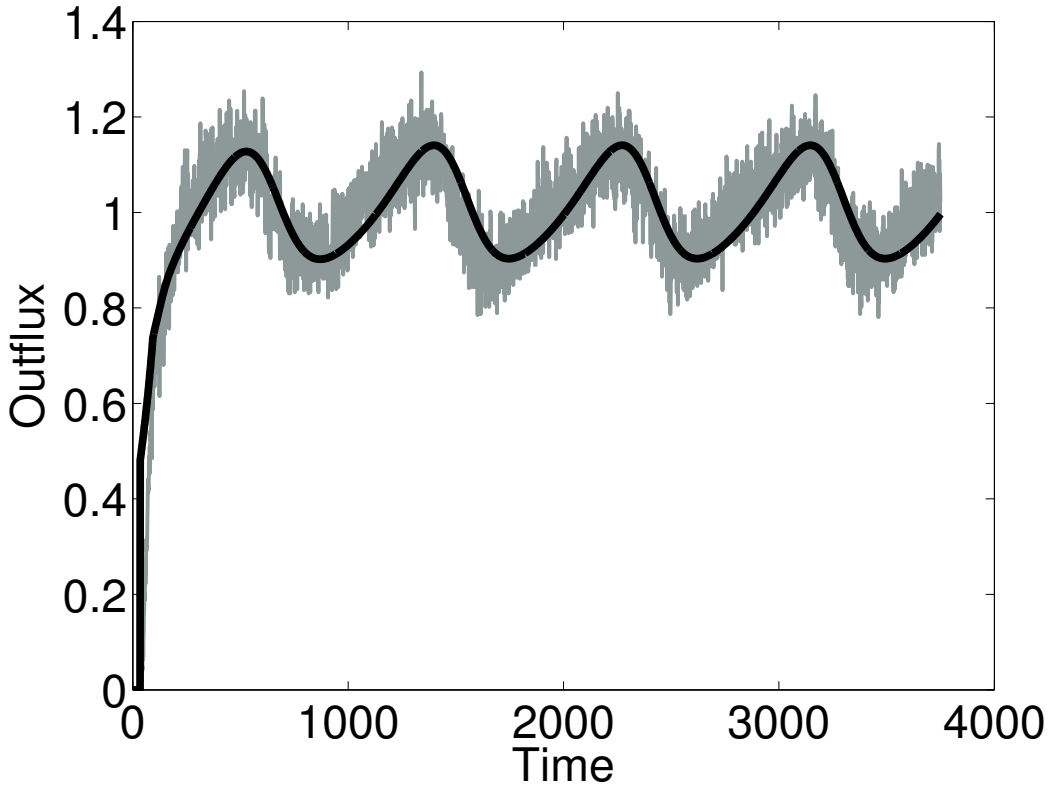


Figure 4.1: Outflux Generated Via a Sinusoidal Influx for Averaging 1,000 DES of a Model of a Semiconductor Factory (Perdaen *et al.* (2008)) (shaded) and a Simulation Based on Equation (4.3) and a Clearing Function Model (black curve).

DES, one arrives at a general flux model $F(x,t) = \rho(x,t)V(W(t))$ for the transport equation (4.3). Figure (4.1) (Lefeber and Armbruster (2011)) compares the outflux of a DES with the influx of a PDE simulation. The noisy (shaded) lies comes from averaging 1,000 DES of a model of a semiconductor factory (Perdaen *et al.* (2008)). The thin curve is generated by solving equation (4.3) with a steady-state velocity model based on relating the steady-state outflux with the corresponding steady-state WIP in the factory model.

These assumptions generate a dramatically improved model for a production process: The model captures the nonlinear dependence of the cycle times on WIP, and it correctly relates the steady-state WIP, cycle time, and out flux. In addition, it

improves the outflow predictions for non-steady-state (i.e. transient) situations compared to static clearing function strategies: Variations in the outflow will be correctly resolved if they result from variations in the local product density that have little effect on the total WIP and hence on the transport velocity. Figure (4.1) shows that, in general, the outflux calculated by the PDE model stays close to the mean outflux provided by the DES.

Observe that the PDE model also allows one to follow the transport of any local WIP portion given by $\rho(x,t)dx$ over time through the factory. Hence, if the observation time interval Δt and the cycle time τ satisfy $\tau \gg \Delta t$, a PDE model enables one to follow the flow of parts through the production unit in contrast to clearing functions or ODE-based models. On the other hand, for short cycle times, say, on the the order of the observation times, the clearing functions based in the total WIP are appropriate. This observation is independent of the velocity model that is used to describe the flow through the factory, i.e. independent of the type of clearing function that is used.

The most problematic feature of the transport model involves the third bullet point above. When the density upstream changes, the velocity on the factory changes nonlocally. This feature makes for a good model for highly reentrant flows with FIFO queue discipline. In this case, a new product entering the factory competes for production capacity with all the products already in the factory, and hence will slow the rate of service of all products, even those nearing completion. For flows that are not reentrant, however, there is little practical justification for why a product that has just arrived to a production system should affect the rate of completion of products much further down the line. This issue on nonlocal velocities is resolved by incorporating a second PDE for the velocity of the products.

Chapter 5

A SECOND ORDER CONTINUUM MODEL

The production planning problem is to determine the production rate of a resource (factory) in the future. This requires an aggregate model for the production flow through the resource. The canonical model for this aggregation is the clearing function model which is based off of the assumption that the local production rate instantaneously adjusts to the one given by the equilibrium relationship between the production rate (flux) and the work in progress (WIP), for example, characterized by queuing theory. This research extends the current theory and modeling for transient clearing functions by introducing our continuum description of the flow of product through the factory based on a second order transport PDE model for the time evolution of the WIP density and the production velocity.

Fundamentally, the production planning problem revolves around identifying the correct influx pattern to a production resource such that the output of said resource over the planning horizon meets some proscribed external demand–or if not met exactly it is within some acceptable tolerance. This problem is complicated by two different major issues: stochasticity and nonlinearity. Stochasticity manifests itself through the uncertainty of the demand and the variation of any demand realization. In addition, variations in the production speed and quality introduce other fundamental stochastic processes. Of course, a production resource can be buffered against many of the more deleterious effects of demand fluctuations though inventory maintenance, yet stochasticity in the production process still requires variable lead time to fill these inventories.

It should be noted that stochasticity is a more fundamental issue than nonlinearity, since the latter is generated by the former via queuing: Nonlinearity is generated by the fact that the variable lead times do not only depend on the stochastic processes that impact production. In fact, the largest contributor to the nonlinearity in a production system is generated by the waiting in queues. Such waiting depends crucially on the amount of material produced concurrently, i.e. the WIP. Specifically, the lead times increase dramatically together with increasing queue length if the flux through the factory closes in on the capacity limit of the production resource. A typical scenario is this: Demand is projected to increase at some time within the current planning horizon. Meeting this demand requires increasing the influx rate into the factory by a lead time earlier than the requested delivery time. However, increasing the start rate will increase the WIP in the factory and, consequently, increase the cycle time. The resulting nonlinear optimization is at the core of the production planning problem.

5.1 Expanding the Transport Model to Second Order

As seen in the previous chapter and will again be seen later in this chapter, the use of a single PDE conservation law, i.e. mass conservation, is insufficient to effectively solve the forward problem. However, mass conservation is a property that is exhibited by queueing systems as the only manner in which the production resource can change its WIP is via the new arrivals or departures from the system. An effective PDE model, therefore should retain this equation. In light of this, an additional PDE is included in the model.

In a series of papers, Armbruster and Ringhofer (2005) and Armbruster *et al.* (2006a,b, 2003, 2006c) developed a kinetic theory of the stochastic transport processes in a factory model. The approach follows turbulence or gas-dynamical modeling of transport processes (Cercignani (1988)). Fundamentally, such an approach is based on a probability density distribution $f(x, \nu, t)$, where

$$f(x, \nu, t) = \Pr\{\xi \in [x, x + dx], \eta \in [v, v + dv], \tau \in [t, t + dt]\} \tag{5.1}$$

describes the probability to find a particle in an x-interval with a speed in a particular ν-interval in a certain time interval. A typical approach to determining the time evolution of such a probability density is to derive equations for the time evolution of its moments relative to the velocity and then to make some closure assumptions to reduce the infinite set of moment equations down to a finite set (Armbruster *et al.* (2003)). The equations for the first two moments are given by

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial [\nu(x,t)\rho(x,t)]}{\partial x} = 0, \tag{5.2}$$

$$\frac{\partial \nu(x,t)}{\partial t} + \nu(x,t)\frac{\partial \nu(x,t)}{\partial x} = 0. \tag{5.3}$$

Together with the steady-state initial values for $\rho(x, 0)$ and $\nu(x, 0)$

$$\rho(x, 0) = \frac{\lambda_1}{\mu - \lambda_1} \tag{5.4}$$

$$\nu(x, 0) = \frac{\lambda_1}{\rho(x, 0)} \tag{5.5}$$

and boundary conditions below imposed on the left boundary, $x = 0$, equations (5.2), and (5.3) are a set of well-posed hyperbolic partial differential equations commonly referred to as the pressureless gas dynamics equations.

Recall that (5.3) is inviscid Burgers' equation. It models the advection of the variable $\nu(x, t)$ that is transported along the characteristics. As a result, once the

initial material has left the domain, the solution is completely determined by the values taken on at the left boundary. This turns out to resolve the issue of nonlocal velocities associated with the one-dimensional model: The time it takes a part ρdx to move through the factory is determined at the time that the part joins the end of the queue. If there is a lot of WIP in front of the new arrival, it will take longer to clear the factory.

The boundary condition for the flux at $x = 0$ is given by the influx, or start, rate $\lambda(t)$,

$$\rho(0,t)\nu(0,t) = \lambda(t) \tag{5.6}$$

The choice of the other boundary condition depends on the stochastic experiment that is described: The expected cycle time, conditioned on the length of the queue needs to be determined. For an $M/M/1$ queue in steady-state, the PASTA (Poisson Arrivals See Time Averages) property suggests that an arriving part will find an average queue length W given by (2.5). Solving (2.5) for λ and substituting this result into (2.7) yields

$$\nu_{ss}(t) \equiv \frac{1}{\tau} = \frac{\mu}{1 + W(t)} \tag{5.7}$$

is the velocity related to the well-known $M/M/1$ clearing function (4.6).

In the general transient case, time averages make no sense any more. Instead, what is of interest is finding the expected time evolution for the movement of parts through the factory given a particular initial state of the system–i.e., the ensemble average, conditioned on the initial WIP.

5.2 Previous Work

Recently, Armbruster *et al.* (2013) discussed a specific discrete-event simulation experiment and showed that we can fit heuristic boundary conditions that allow us to reproduce the ensemble averages of the experiments. The experiment in question is that suggested by Missbauer (2011) in his work on transient clearing functions: Consider an $M/M/1$ queue with a production rate of $\mu = 1/day$, and determine the ensemble average of the total number of parts in the production unit over this five day period. Calling the initial queue $W(0)$, the ensemble average for the total number of parts becomes the sum of the initial queue and the cumulative influx $\lambda(t)$:

$$L = W(0) + \int_0^5 \lambda(t)dt. \tag{5.8}$$

Our heuristic is based on two regimes:
(1) If $\lambda(t) < \mu$, then we expect that any initial WIP distribution decay exponentially fast to the WIP distribution associated with the stead-state related to the arrival rate. This is know for Markov processes as the *mixing time* (Levin *et al.* (2009)). Hence, the boundary condition is determined by the solution to an ordinary differential equation,

$$\frac{d\nu(0,t)}{dt} = -\sigma(\nu(0,t) - \nu_{ss}(t)) = -\sigma\left(\nu(0,t) - \frac{\mu}{1+W(t)}\right), \tag{5.9}$$

where the decay constant σ will be determined experimentally. (2) If $\lambda(t) > \mu$, the queue length will become unbounded. Assuming that over the course of the observation interval of five cycles the queue never becomes empty, and therefore the machine is never starved, the cycle time at arrival of a part at a queue length of $W(t)$ will become just $\tau = W(t)/\mu$. For arrival rates only slightly larger than the

production rate and for small $W(0)$, this is not always the true. We found that with a velocity of

$$\nu(t) = \frac{\mu}{0.5 + W(t)} \tag{5.10}$$

we obtain good agreement between PDE simulations and ensemble averaged DES.

Hence the full boundary condition for (5.3) becomes

$$\nu(0,t) = \frac{\mu}{0.5 + W(t)} \qquad \text{for } \lambda \geq \mu, \tag{5.11}$$

$$\frac{d\nu(0,t)}{dt} = -\sigma\left(\nu(0,t) - \frac{\mu}{1 + W(t)}\right) \qquad \text{for } \lambda < \mu, \tag{5.12}$$

$$\nu(0,0) = \frac{\mu}{0.5 + W(0)}. \tag{5.13}$$

The last equation (5.13) describes the initial condition for the ordinary differential equation. It is based on the assumption of a deterministic initial condition, i.e., the initial WIP is exactly known, and hence the ensemble average will be mostly affected by the stochasticity of the machine process and little affected by the stochasticity of the arrival process. Figure 5.1 shows comparisons of DES and PDE simulations of Missbauer's experiments for different influxes and initial WIP.

While the model given above shows that the PDE aggregates the flow through the production resource with reasonable accuracy, when we view the outflux profile for this model we can see that this model falls short. Figure (5.2) shows a ramp-down transition of $\lambda_1 = 0.5 \to 0.2 = \lambda_2$ that is representative of the outflux patterns observed when running our initial model above. We obtained similar results for the ramp-up transition $\lambda_1 = 0.2 \to 0.5 = \lambda_2$. Clearly this model does not describe the

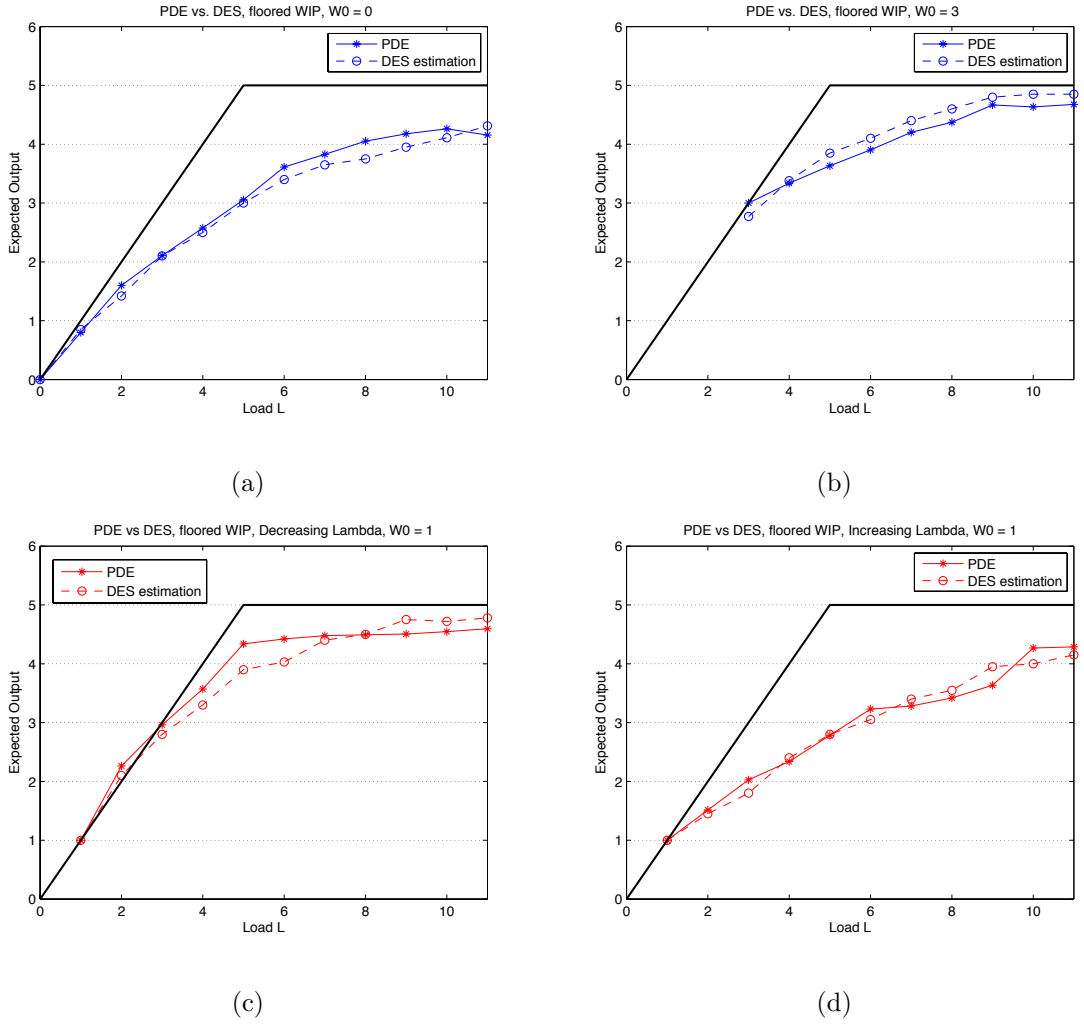


Figure 5.1: Outflux Over Five Time Intervals as a Function of the Total Expected Load for DES and for the PDE Model (5.2, 5.3) with Boundary Conditions (5.6) and (5.11 - 5.13). a) Constant Influx and Initial WIP of $W(0) = 0$, b) Constant Influx and Initial WIP $W(0) = 3$, c) Decreasing Influx for $W(0) = 1$, d) Increasing Influx for $W(0) = 1$

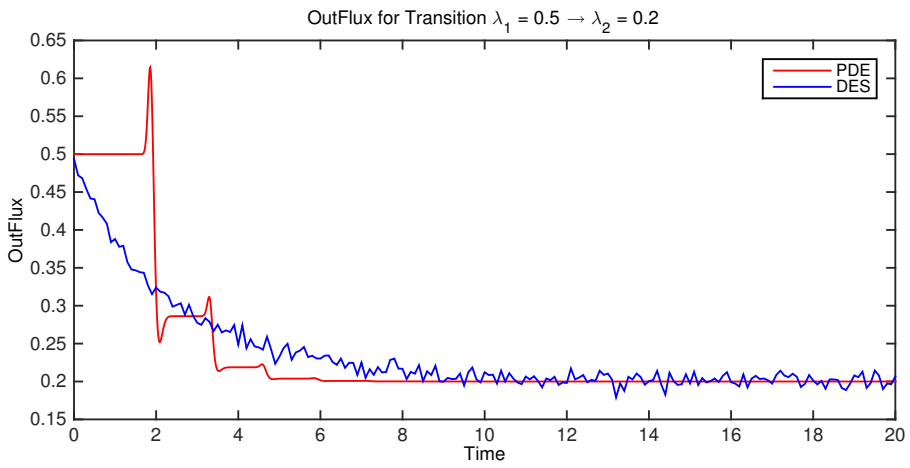


Figure 5.2: Initial Model DES and PDE Outflux for Stepwise Transition $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.3$.

evolution of the flow accurately.

5.3 New Model

With these results from our initial attempt, a simpler model was examined in that the the steady-state boundary condition (5.7)–extended to include the point $\nu(0, 0)$–rather than the ODE boundary condition (5.9) was used. In addition, the exponential scenario will be looked at first to get an idea of how well the new model performs as the plateaux which are due to the propagating discontinuity of a stepwise influx can be avoided.

Examining the transition from $\lambda_1 = 0.7$ to $\lambda_2 = 0.3$, Figure 5.3 shows that the new continuum model performs moderately well. In the long term, the PDE converges to the correct steady-state, which is desirable. However, the transition time is significantly faster than what is observed in the DES. This is compounded by the fact that the DES shows an immediate movement towards the new steady-state whereas the PDE maintains a level output at the original steady-state level of $\lambda_1 = 0.7$. This

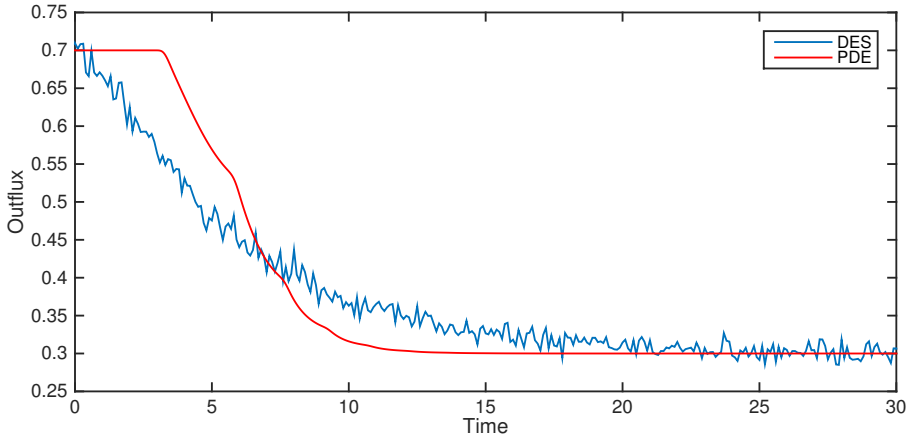


Figure 5.3: DES and PDE Outflux for Exponential Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$.

constant initial outflux in the PDE is a byproduct of the hyperbolicity of the continuum model.

Looking at the transition in the opposite direction, that is, from $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$, the accuracy of this PDE model is far better. As Figure 5.4 illustrates, the convergence to the steady-state at $\lambda_2 = 0.7$ in the PDE solution is very close to that of the DES. Again, though, one can see the characteristic constant outflux for the first few time cycles that is due to the initial WIP travelling with the steady-state velocity for $\lambda_1 = 0.3$.

Clearly the lack of symmetry between increasing transitions and decreasing transitions must be due to the dynamics of the initial WIP in the system. In the upwards transition the total WIP is increasing, so that the the only influence of the initial WIP is the delay. In the decreasing transition, though, there is excess WIP that must be drained in the transient phase. This means that the dynamics of the initial WIP has greater impact on the output for decreasing transitions than for increasing ones.

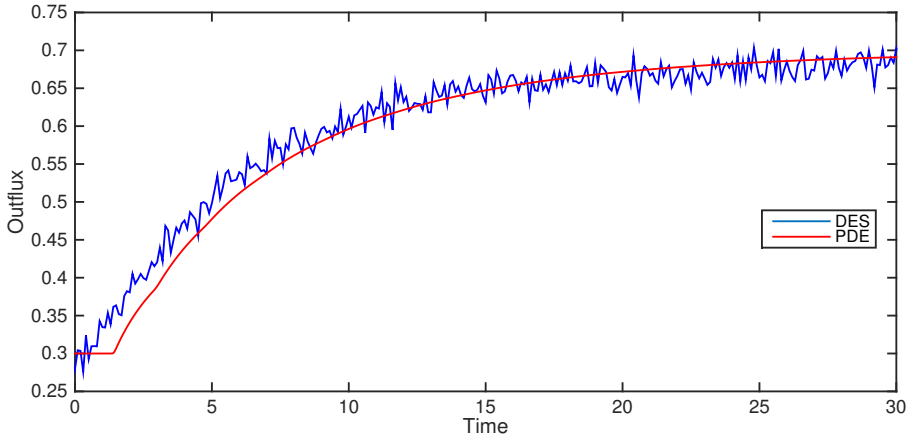


Figure 5.4: DES and PDE Outflux for Exponential Transition $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$.

Returning to the stepwise input pattern and analyzing its solution yields a better understanding of the dynamics of the WIP. Using the same transition values as in the exponential case, Figure 5.5 and Figure 5.6 show the results of the experiment with a stepwise input pattern. Again, one can see the convergence to steady-state for the PDE is faster than the DES. Also, the constant outflux for the first few time steps is present as in the exponential case. Unfortunately, for the stepwise scenario there are the plateaux due to the discontinuity. Since these plateaux are not represented in the DES, this suggests that this completely hyperbolic model is insufficient to incorporate the dynamics of the WIP.

Before moving on, it is beneficial to compare the above results with those obtained from the first order model provided in chapter 4. Figures 5.7 and 5.8 illustrate the results of the experiments for stepwise and exponential transitions for the same λ values as above. The most glaring feature of the this PDE solution is that it generates large outflux spikes in the opposite direction of the transition over the same time period where the second order model has constant outflux. Also, in most cases, there are corners (but no plateaux) found in place of the plateaux seen in the second order

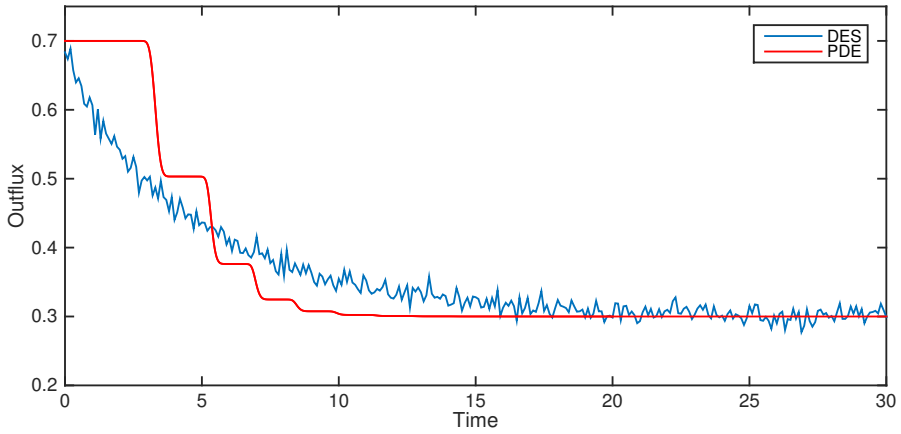


Figure 5.5: DES and PDE Outflux for Stepwise Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$.

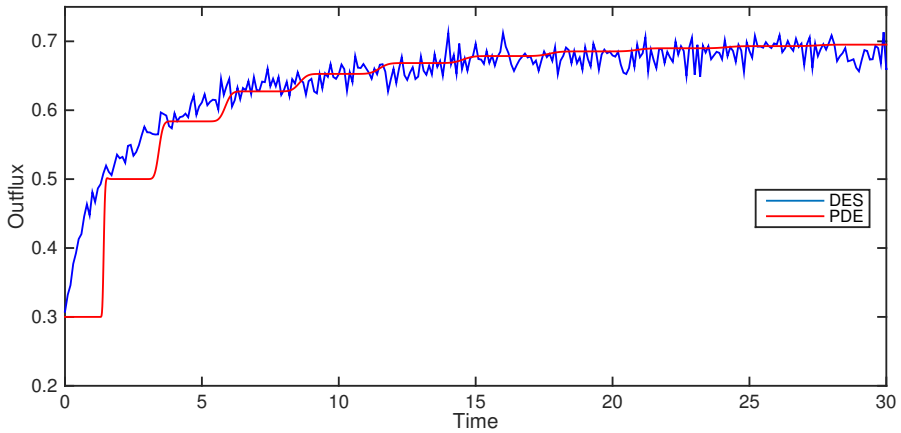
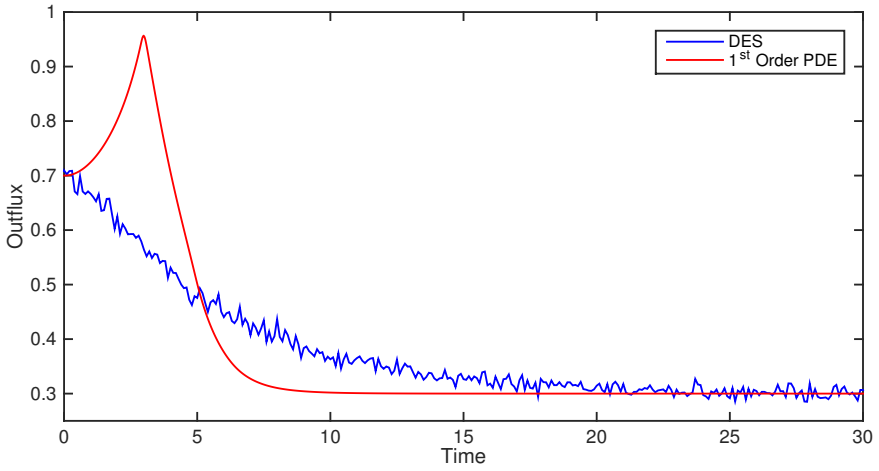


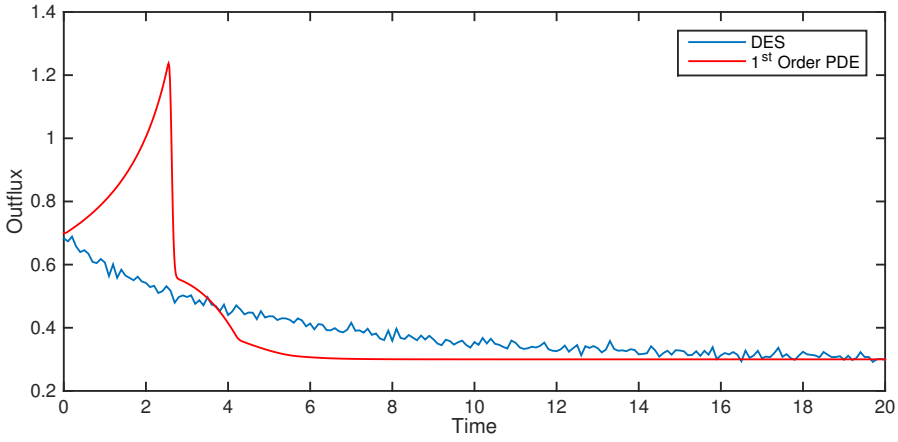
Figure 5.6: DES and PDE Outflux for Stepwise Transition $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$.

model. Both of these aspects are due to the nonlocalized velocity condition. As mass enters/leaves the system the velocity changes as it does in the second order model, but it does so at every point x identically and instantaneously. Therefore, the outflux being the product of the density ρ and the velocity ν will see an increase/decrease based on the increase/decrease of ν since the ρ advects with finite speed.

While the specific transitions $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$ and $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$ were the focus here, the other experimantal transitions had similar results. The complete list of experimental exponential/step-wise transitions is contained in Table 5.1.

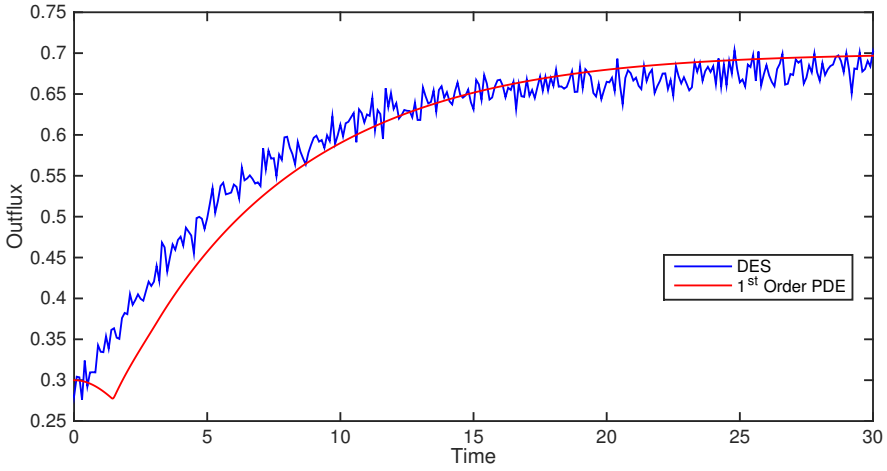


(a) Exponential transition.

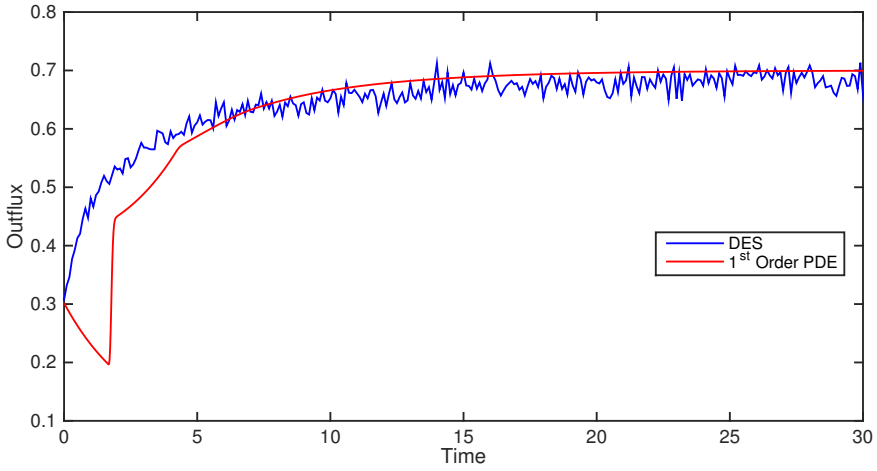


(b) Stepwise transition.

Figure 5.7: 1^{st} Order PDE Solution $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$.



(a) Exponential transition.



(b) Stepwise transition.

Figure 5.8: 1^{st} Order PDE Solution for $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$.

5-Series Transitions	
λ_1	λ_2
2	5
3	5
4	5
5	2
5	4

7-Series Transitions	
λ_1	λ_2
3	7
5	7
6	7
7	3
7	5
7	6

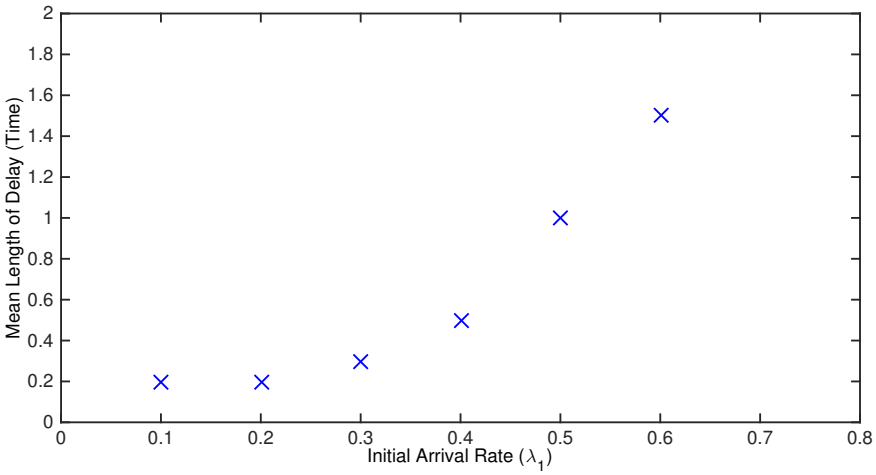
9-Series Transitions	
λ_1	λ_2
1	9
3	9
5	9
7	9
9	1
9	3
9	5
9	7

Table 5.1: Table of Transitions Subject to Experimentation.

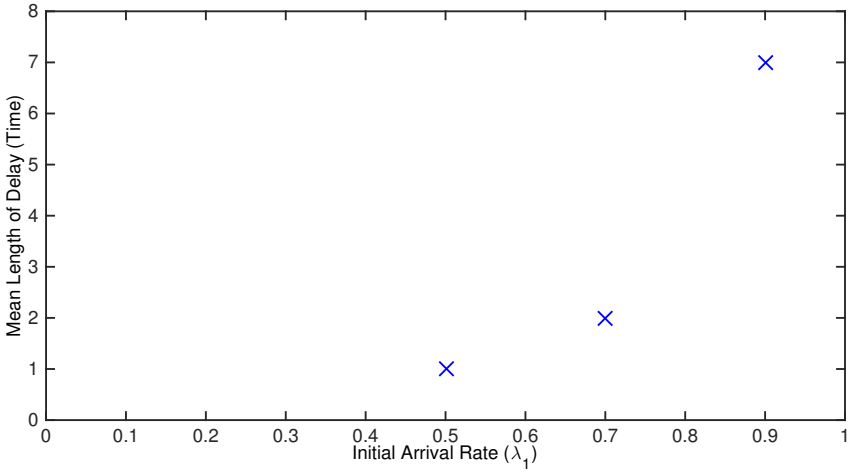
5.4 Adding Diffusion to the Model

In both the ramp-up and ramp-down examples discussed above the PDE outflux spends more time at the initial steady-state levels then the DES suggests that it should. In the ramp-down scenario, the delay in the movement of the outflux implies that there is too much material moving through the resource at too high a velocity. The ramp-up case is the reverse, where not enough of the material is flowing through the resource fast enough. To rectify this, a diffusion term is added to the mass conservation PDE that will speed up or slow down the flow through the resource depending on the situation–ramp-up or ramp-down. This feature of the PDE model is not restricted to the specific transitions presented in the last section. Figure 5.9 shows that this delay is present to varying degrees for all the transitions given in Table 5.1.

Diffusion is incoporated into the model by adding a second order term of the



(a) Ramp-up delay time.



(b) Ramp-down delay time.

Figure 5.9: Mean Time Delay Before PDE Moves from Initial Steady-State

density in the mass conservation PDE (5.2). This term has the general form

$$D\rho_{xx}(x,t) \tag{5.14}$$

with diffusivity constant $D > 0$. Hence, the new PDE pair becomes

$$\rho_t(x,t) + [\nu(x,t)\rho(x,t)]_x = D\rho_{xx}(x,t), \tag{5.15}$$

$$\nu_t(x,t) + \nu(x,t)\nu_x(x,t) = 0. \tag{5.16}$$

The IC and BC remain (5.4), (5.5) and (5.6), (5.7), respectively, and the addition of the following one provides a well-posed problem.

$$\rho_x(0,t) = 0 \tag{5.17}$$

In flux conservative form, the new mass conservation PDE is

$$\rho_t(x,t) + [\nu(x,t)\rho(x,t) - D\rho_x(x,t)]_x = 0 \tag{5.18}$$

The diffusive flux term $-D\rho_x$ moves mass along the negative gradient of the concentration of the material at a point x at time $t > 0$. In the ramp-down case the influx on the left boundary and this diffusion serves to move material to the left since a decreasing influx will predominantly affect the density level by reducing it given that the velocity is defined as a function of the total WIP and, therefore, lags behind changes in the density. Similarly, in the ramp-up case, diffusion advances material to the right, thus causing more material to exit the system sooner.

However, when attempting to implement this advection-diffusion model to the above transitions the results are mixed. For the stepwise transition $\lambda_1 = 0.7 \to \lambda_2 = 0.3$ the addition of diffusion eliminates the plateaux as desired. It also reduces the length of time that the outflux remains at the initial steady-state. The downside is

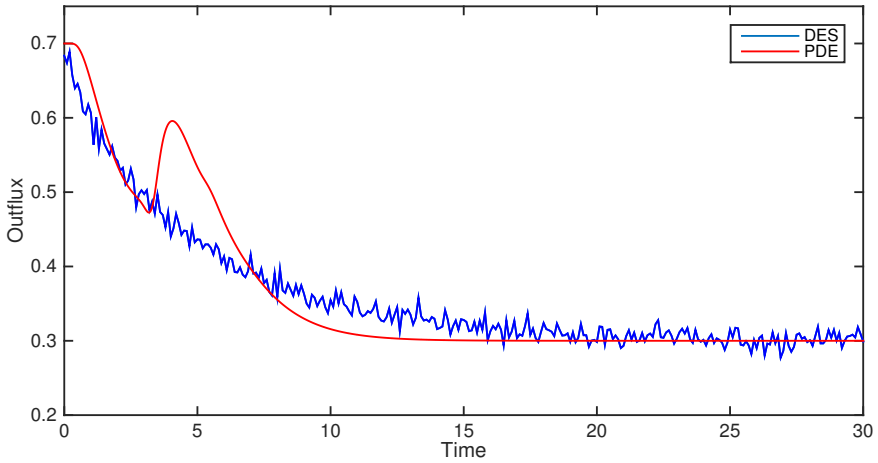


Figure 5.10: Solution for (5.15) (5.16)–Stepwise Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$ with Diffusion Coefficient 0.10

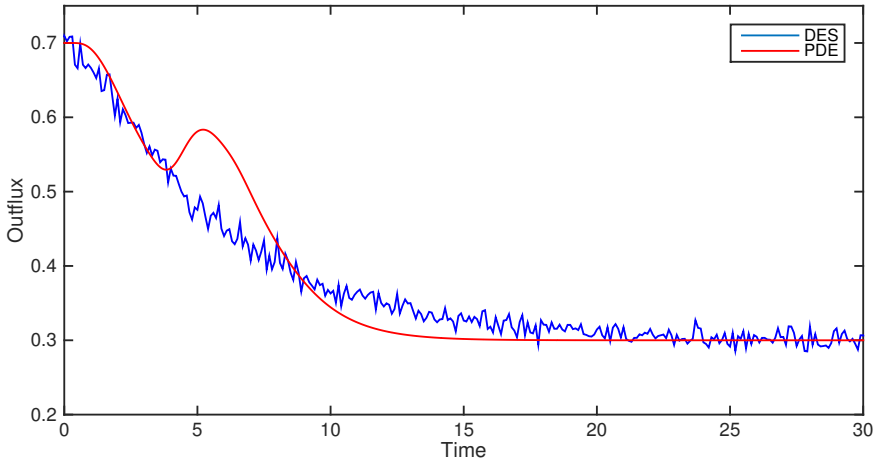


Figure 5.11: Solution for (5.15) (5.16)–Exponential Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$ with Diffusion Coefficient 0.10

clear in Fig 5.10 that the addition of diffusion creates a delayed wave-like behavior. Similar results are obtained for the exponential transition with the same λ values as seen in Fig 5.11 although the severity of the this behavior is lessened.

The increasing transitions are generally much better because of the lesser impact of the initial WIP. For the stepwise transition with a diffusion coefficient of 0.10, the solution in Figure 5.12 does not produce as nice a result as the exponential in Figure

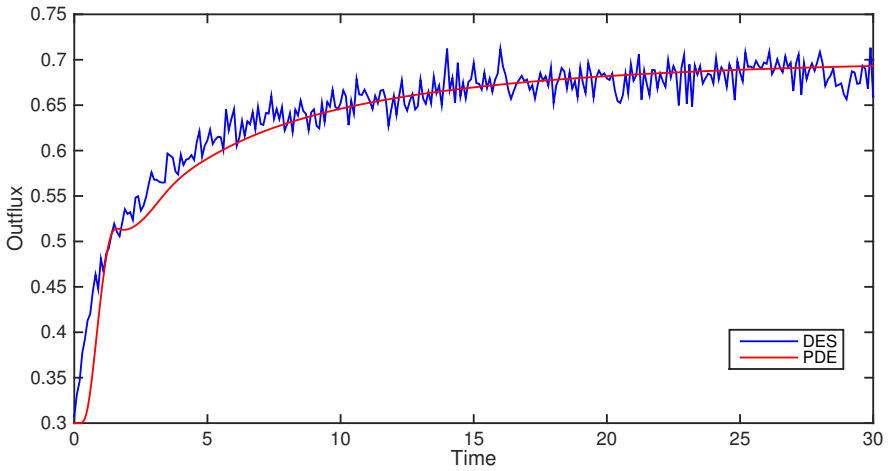


Figure 5.12: Solution for (5.15) (5.16)– Stepwise Transition $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$ with Diffusion Coefficient 0.10

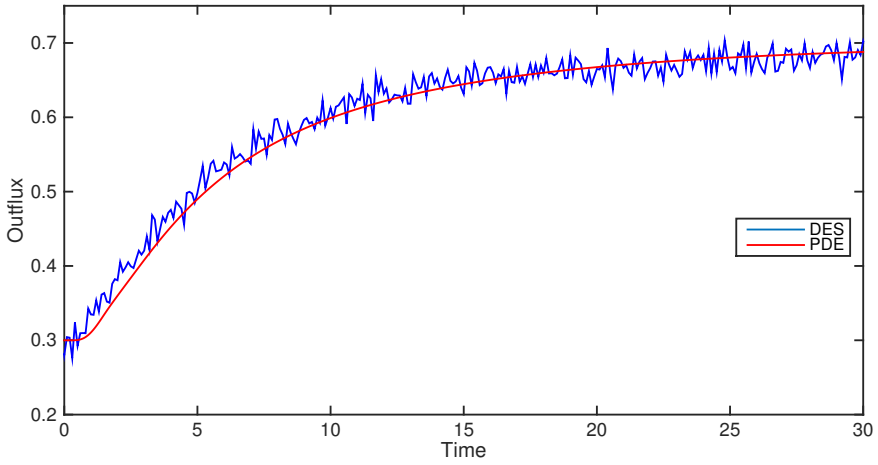


Figure 5.13: Solution for (5.15) (5.16)– Exponential Transition $\lambda_1 = 0.3 \rightarrow \lambda_2 = 0.7$ with Diffusion Coefficient 0.10

5.13 due to the abrupt change in the outflux for the stepwise transition around $t = 2$.

This wave-like behavior due to the diffusion is not shared by all transitions. Figures (5.14) and (5.15) show the outflux of the PDE for the increasing and decreasing exponential transitions of $\lambda_1 = 0.2 \rightarrow \lambda_2 = 0.5$ and $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.2$, respectively. From the plots, one can see that the diffusion term influences the outflux in the expected manner. The delay in moving off of the initial steady-state outflux has

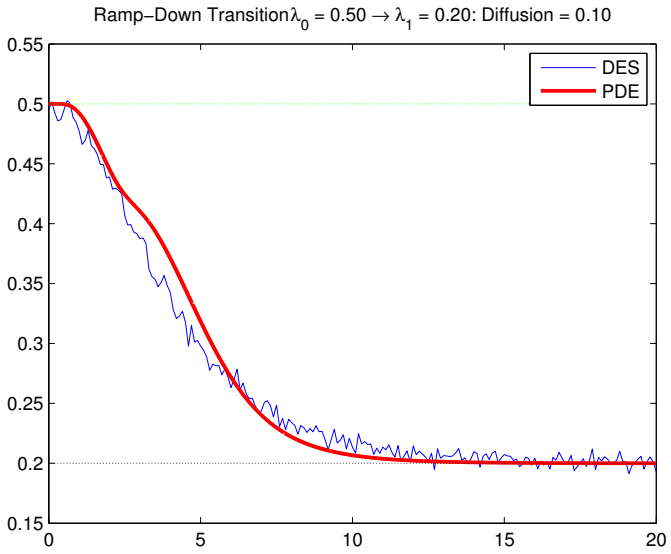


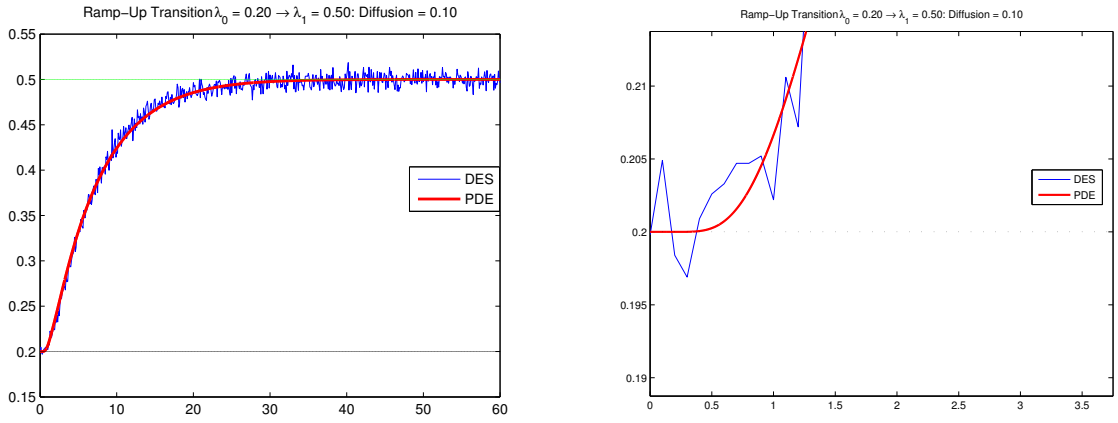
Figure 5.14: Solution for (5.15) (5.16)–Exponential Transition $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.2$ with Diffusion Coefficient 0.10.

been reduced.

5.5 Long-Term Time-Varying Behavior

The PDE and the DES for the exponential and stepwise input pattern dealt entirely with the transient behavior of the queueing system. In the long term as the previous scenarios illustrate, the PDE converges to the same steady-state as the DES. This is surely a desirable feature in the model. However, consistency with long term DES steady-states is not the sole interest. Also of interest is how well the PDE model represents long term queueing behavior that is not steady-state. By experimenting with cyclic input patterns, the efficacy of the second order model can be evaluated for time varying influxes over time frames for which the transient behavior has passed.

In Figure 5.16 one can see the general characteristics in the resulting outflux due to cyclic input. For this example, the input pattern has the formulation in Eq(3.3)



(a) Exponential Transition with Diffusion Coefficient 0.10.

(b) Exponential Transition with Diffusion Coefficient 0.10, zoomed in.

Figure 5.15: Solution for (5.15) (5.16)–Exponential Transition $\lambda_1 = 0.2 \rightarrow \lambda_2 = 0.5$ with Diffusion Coefficient 0.10.

with $\lambda_1 = 0.7$, $\lambda_2 = 0.3$, and $C = 5$. The second order continuum model produces a behavior that is quite similar to that given by the DES in both the short term, transient phase as well as the long term steady phase. Both plots appear to have the same transient relaxation as they settle into a stable oscillatory motion about the mean of 0.5. Additionally, the amplitudes are approximately the same with, perhaps, the PDE being slightly larger. Lastly, the period of the DES and the PDE are the same and correspond to the period of the input function. Thus, there appears to be no frequency modulation on the part of the system.

Issues still remain, however. Again it can be seen that the initial outflux of the PDE is constant for the first few time steps just as in the previous scenarios - although, one should not have expected anything different regarding this input type. One also can observe a phase shift in the PDE relative to the DES. Finally, there is a clear narrowing of the amplitude as the input pattern is processed through the system. Both of these features are examined in turn.

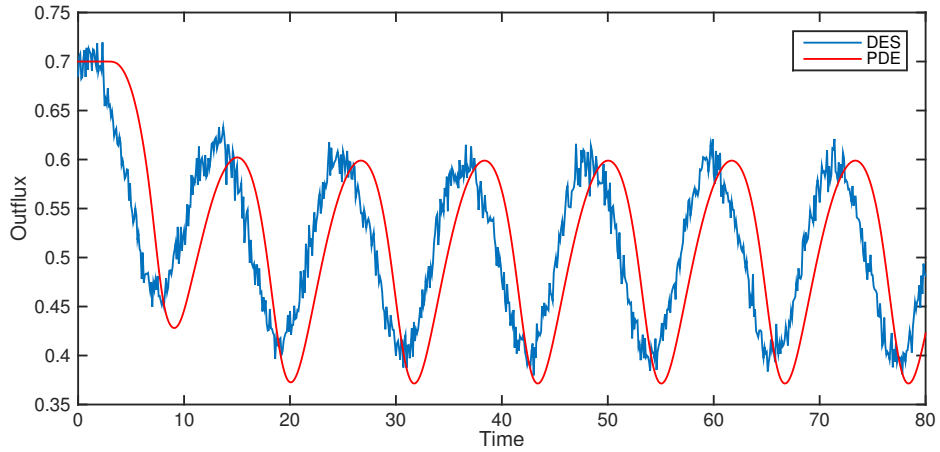


Figure 5.16: DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.7] and $C = 5$.

The phase shift is the lion's share of the discrepancy between the DES and the PDE with the PDE lagging behind the DES in time. This is due to the initial constant outflux that has presented itself in all the PDE results so far. As Figure 5.17 illustrates, when this constant outflux is accounted for, the PDE and DES are a good match. In practice, the phase shift does not pose any particular threat to the efficacy of the PDE model to solve the forward problem for this type of influx function. This is because one need only wait for a few cycles to determine the correct phase and then adjust the production flow for the discrepancy.

As mentioned earlier, the attenuation of the amplitude is shared by the DES and the PDE. When the influx and the outflux for the DES are directly compared using the same λ parameters above-as in Figure 5.18-one can see the scale of the narrowing is not insignificant. Moreover, this reduced outflux amplitude is evident in all the other experiments that were run. Performing the Fourier analysis for the cycle with $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$ across the eight different periods $C = 10, 7, 5, 3, 2, 1.8, 1, 0.5$

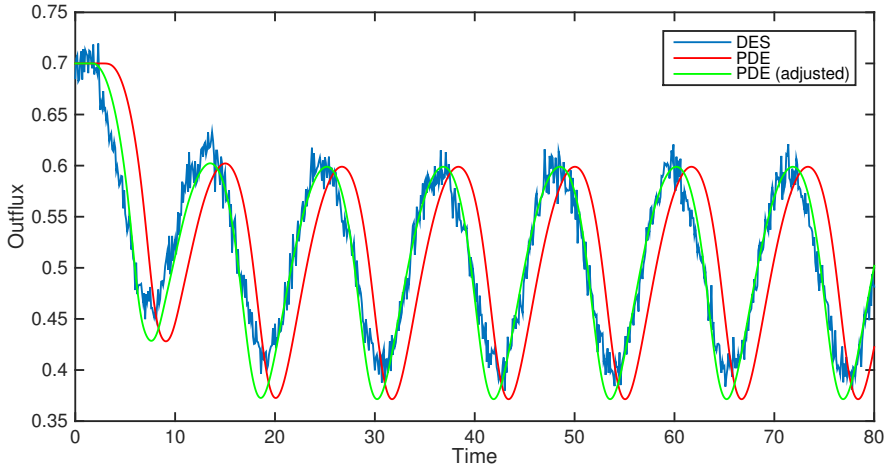


Figure 5.17: Reprint of Figure 5.16 with Phase Adjusted Solution (green).

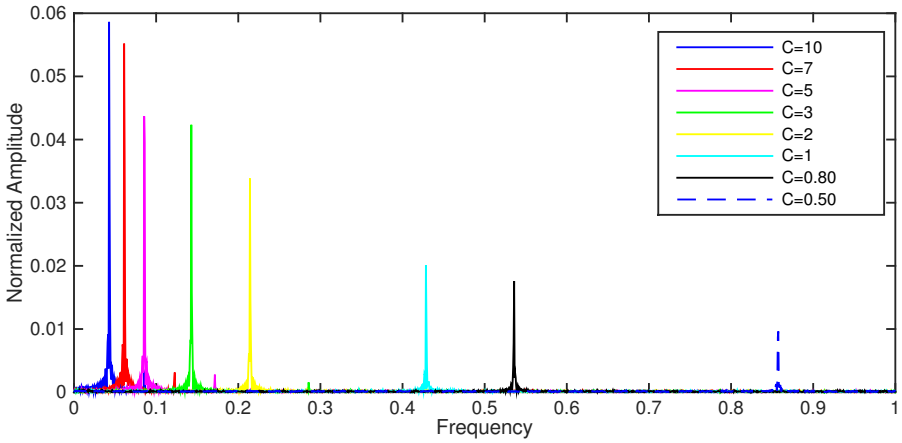


Figure 5.18: Normalized Amplitude Periodogram for Outflux Derived from Cyclic Influx with Range [0.3, 0.7]

reveals a pattern of amplitude narrowing of varying degree dependent on the period of the influx function. The normalized amplitude periodogram in Figure 5.18 illustrates this fact. As the period shortens, the magnitude of the attenuation increases. The periodogram also confirms that there is only one mode for each value of C thus eliminating the possibility of frequency splitting.

For high frequency forcing ($C < 2$), the results are interesting and problematic.

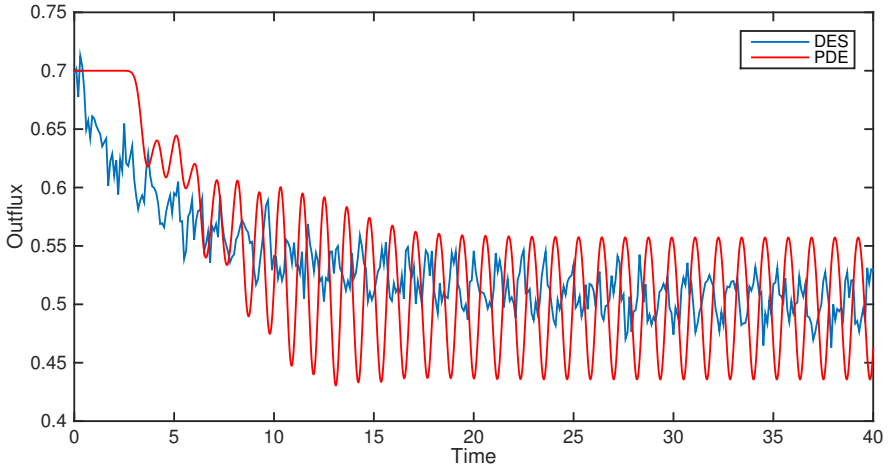
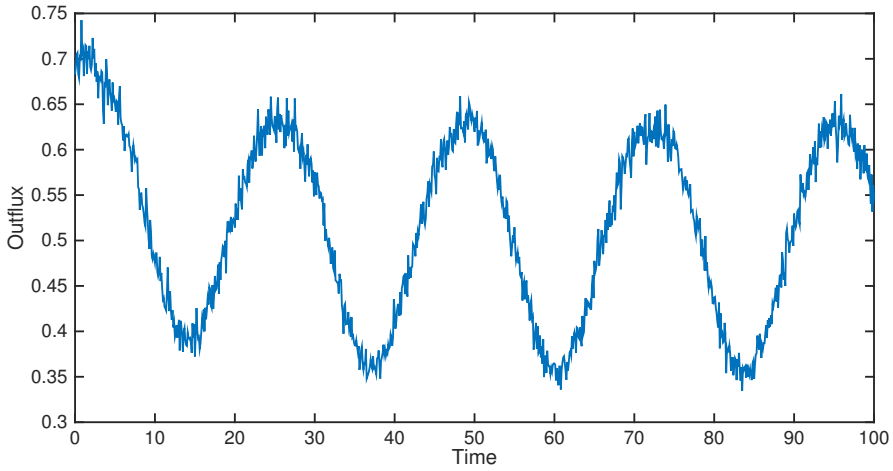


Figure 5.19: DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.7] and $C = 0.5$.

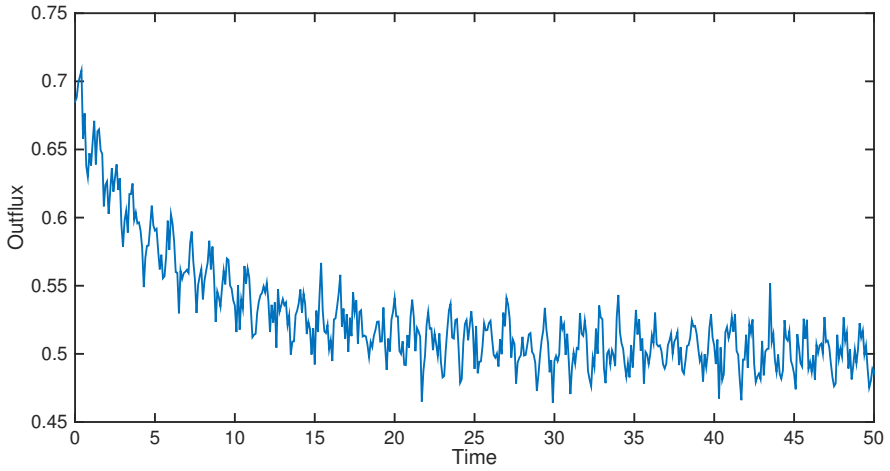
Examination of the outflux in Figure 5.19 shows the expected narrowing of the amplitude. However, the corresponding PDE does not have a matching amplitude. The figure also shows that the oscillations of the DES are becoming less well defined. Further inspection of the DES for $C = 0.5$ given in Figure 5.20(b) seems to suggest a breaking down of the wave pattern by the system. Contrasted with Figure 5.20(a), this phenomenon is quite severe. The issue here is the low signal-to-noise ratio resulting from high frequency input patterns, but because it is stochastic in nature, it cannot be accounted for using PDEs. [1]

The DES and PDE have a matching attenuation of the amplitude as illustrated in Figure 5.16, but this is not characteristic across all the frequencies in the experiments as Figure 5.19 has shown. For the frequencies corresponding to C values of {2, 3, 5, 7, 10} there is an excellent matchup with the amplitude for the DES and the PDE Figure 5.21. Around values of $C = 0.8, 1, 1, 8$ a distinct divergence between the two

[1]It should be pointed out that the decreasing signal-to-noise ratio is not the result of the simulation techniques used as the DES input patterns do not exhibit any change in the signal-to-noise ratio regardless of the value of C.



(a) $C = 10$.



(b) $C = 0.5$.

Figure 5.20: DES Outflux for Selected C Values for Cyclic Influx with Range [0.3, 0.7]

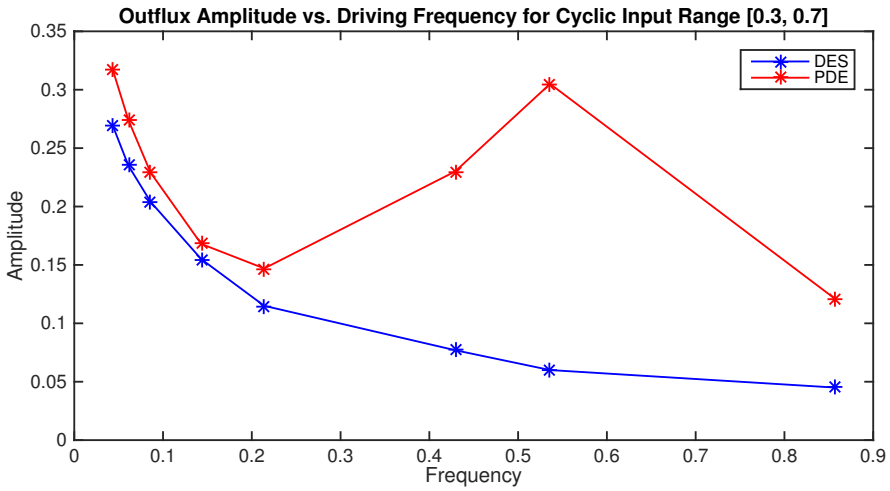


Figure 5.21: Normalized Amplitude Versus Influx Frequency for DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.7]

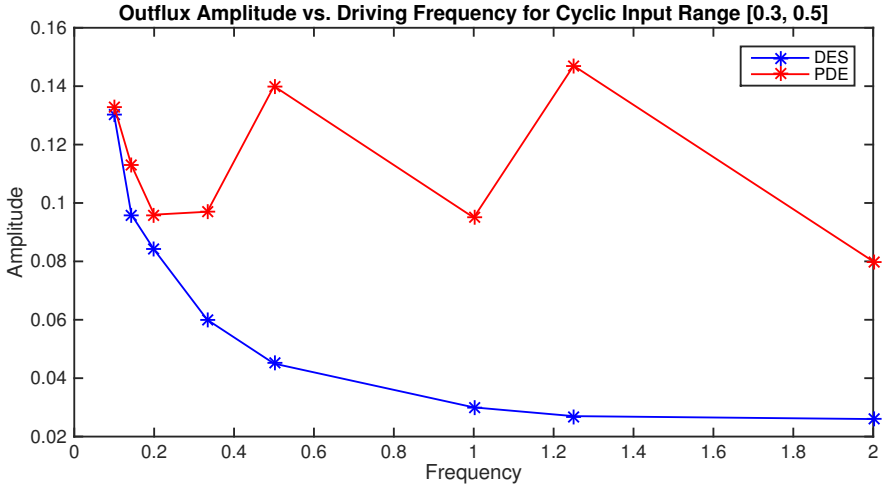


Figure 5.22: Normalized Amplitude Versus Influx Frequency for DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.5]

arises. The cause for this separation is because the PDE has a resonance frequency at 0.5. As well, this behavior is found in the other cyclic input pattern experiments as depicted in Figures 5.22 and 5.21). Unfortunately, this cannot be resolved at this time and is where the continuum model fails to solve the forward problem.

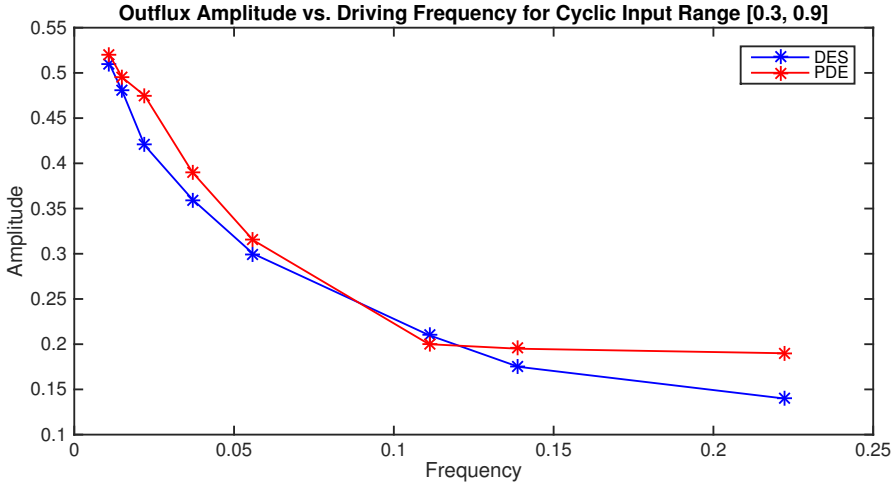


Figure 5.23: Normalized Amplitude Versus Influx Frequency for DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.9]

Chapter 6

CONCLUSIONS AND FUTURE WORK

Chapter 1 began this research with a description of the two fundamental production planning problems, namely, the forward problem and the backward problem. In the forward problem one is given a known set of inputs or input (influx) pattern and the solution to said problem is the determination of the corresponding set of output or the output (outflux) pattern. Some of the more popular approaches in the literature, specifically, linear programming with exogeneous fixed lead times and clearing function models, were introduced and discussed. The advantages and limitations that these models have in solving the forward problem were highlighted as well. This work seeks to overcome the limitations of these models by approaching the forward problem with a continuum model.

In chapter 4, the basics of the continuum models were introduced with a specific focus on the ODE fluid models and first-order transport PDE models. The prior established shortcomings of the LP and clearing function models, that being the inability of the former model to capture the nonlinear relationships between workload and cycle time and the latter's complicated approaches to incorporating the the delay between influx and outflux and input pattern distribution, are largely overcome by the aggregating of the stochastic flow through the factory. This is accomplished in two ways: 1) averaging over time (or ensemble) the desired measures of effectiveness of the system, thus converting the stochastic process into a deterministic one; 2) converting the WIP into a real-valued quantity from the integer value held by the other models.

The second order continuum model introduced in this research expands on the success of its PDE predecessors by incorporating an equation for the velocity of the material flow. In chapter 5, it was shown that the first order PDE model does not reflect accurately the transient behavior of the $M/M/1$ system with its large outflow spikes and precipitous drops. The use of non-local velocity constraints is the root cause of this behavior and the second equation for velocity takes care of this. Both increasing and decreasing transitions show improved accuracy in modeling the queueing system at hand with the improvements in scenarios involving low utilization levels (e.g. $\lambda/\mu \leq 0.5$) and increasing transitions being the most noteworthy.

The downside to the second order model is that both shock and rarefaction waves may be present in the solution. Rarefaction waves appear in the increasing transitions and generally help in the performance of the model. Shock waves, on the other hand, are not desirable and result from decreasing transitions. The addition of a diffusion term improved the performance of the PDE model by eliminating the plateaux caused by propagating discontinuities as in the stepwise experiments and reducing the duration of the initial constant outflux profile. The cost for this addition is that the remaining strong gradients generate a buildup of material on the interior of the domain manifesting in a single, wave-like pattern passing through the system. Unfortunately, while increasing the diffusion strength would likely remove this new behavior it would also weaken the hyperbolicity of the model. Substantial loss of hyperbolicity is a far greater concern than that of the merged shock.

The second order model performs quite well for more complicated input patterns, as well. In the cyclic scenario, it was shown that the model is very accurate for

longer period oscillations - those on the order of twice the average cycle time or greater. When one accounts for the initial constant outflux duration, the DES and PDE model are very much in tune. Both exhibit an attenuation of their output pattern amplitudes relative to the input patterns' amplitudes with the magnitude of this narrowing being a function of the driving frequency.

The cyclic experiments showcased two new issues unique to this input pattern type. First, as the frequency of the influx increases the signal-to-noise ratio falls. While this most likely contributes in part to the overall attenuation of the amplitude, it is in the high frequency regimes ($C < 2$) that problems with the DES start to arise. At these frequencies, the signal-to-noise ratio is so low that the oscillatory pattern begins to break down. This is a problem involving the stochasticity of the $M/M/1$ system and cannot be accommodated using PDEs. Second, the PDE exhibits resonances at the higher frequencies which serve to create much larger outflux amplitudes relative to their corresponding DES outflux amplitudes. This is endemic of the hyperbolicity of the PDE model and as such is not apparent in the DES.

Overall, this second order model derived from kinetic theory performs remarkably well in solving the forward problem set up in this paper. There is surely room for improvement, yet, the results of the core model suggest that the potential of this model is significant.

6.1 Future Work

The promising results for this second order, hyperbolic continuum model suggest further avenues of research that may prove fruitful.

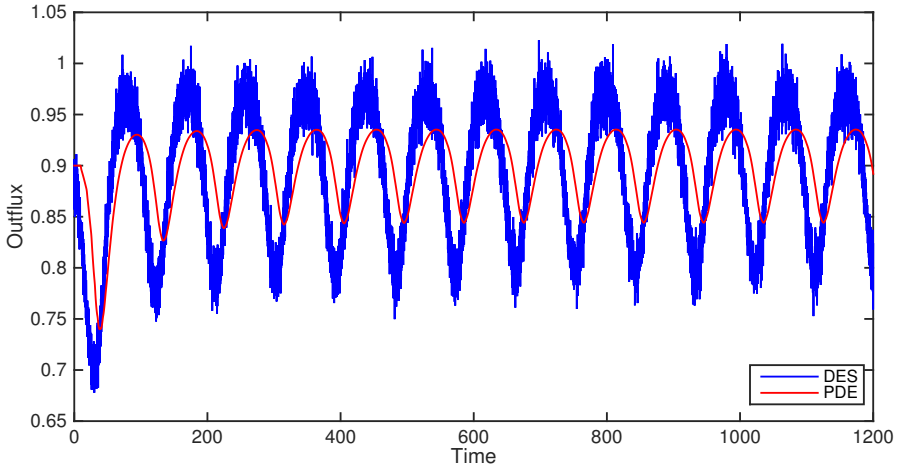


Figure 6.1: DES and PDE Outflux Derived from Overloaded Cyclic Influx with Range [0.6, 1.2]

6.1.1 Overload Experiments

Of the three scenarios, exponential, stepwise, and cyclic, the cyclic scenario affords one a special experiment. Recall from chapter 2, that for an arrival rate λ the queueing system will not attain a steady state if $\lambda \geq \mu$, rather the queue will continue to grow without bound. However, this was based on λ remaining constant as $t \to \infty$. If λ is a function of t, then it is possible to have a stable queueing system even if $\lambda(t) \geq \mu$ as long as the mean value of $\lambda(t) < \mu$.

In Figure 6.1 one can see the results of the only experiment that was run investigating the overload scenario. Starting from a steady-state $\lambda_1 = 0.9$, $\lambda(t)$ oscillates about this value between a maximum of $\lambda_{max} = 1.2$ and minimum of $\lambda_{min} = 0.6$. The transient seems to match reasonably well, however in the long term there is the unusual situation where the PDE amplitude attenuates to a greater degree than the DES. It remains unclear as to the reason for this. This issue is compounded by the fact that the PDE and DES do not appear to match in their amplitudes for any

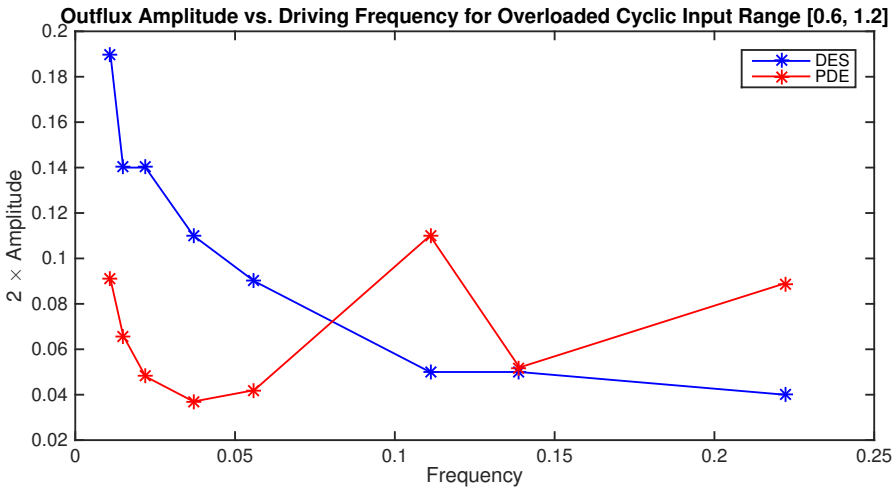


Figure 6.2: 2 × Normalized Amplitude Versus Influx Frequency for DES and PDE Outflux Derived from Cyclic Influx with Range [0.3, 0.7]

of the subject experimental frequencies as shown in Fig 6.2 even in the high C/low frequency regime.

6.1.2 Understanding the Diffusion Terms

In the latter part of Chapter 5, a diffusion term was introduced to this continuum model which gave the system (5.15), (5.16) with initial conditions and boundary condition (5.4), (5.5) and (5.6), (5.7).

The original impetus for this was the delay encountered in the PDE model when moving off the original steady-state level λ_1. Adding a diffusion term of the form $D\rho_{xx}$ for constant $D > 0$ enabled the adjustment of the outflux profile for the PDE model so that it fell more in line with the DES data. However, this addition while working well for the transitions $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.2$ and $\lambda_1 = 0.2 \rightarrow \lambda_2 = 0.5$ performs poorly for transitions from higher steady-state levels of λ_1. For example, the transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$, illustrated in Figure 6.3, displays an unexpected crest during the transition time frame that was not identified in the previous, smaller

transitions. The height of the crest is dependent on the magnitude of the diffusion coefficient and is believed to be the consequence of the shock smoothing. As discussed earlier, increasing the diffusion constant would smooth the shock further, but at the expense of losing hyperbolicity. In the increasing transitions, this is not an issue, since the diffusion acts in a favorable manner for the rarefaction wave that is present. Of course, this effect can be mitigated by implementing smaller diffusion coefficients, yet this course then reduces the effect of the correction for which the diffusion term was introduced.

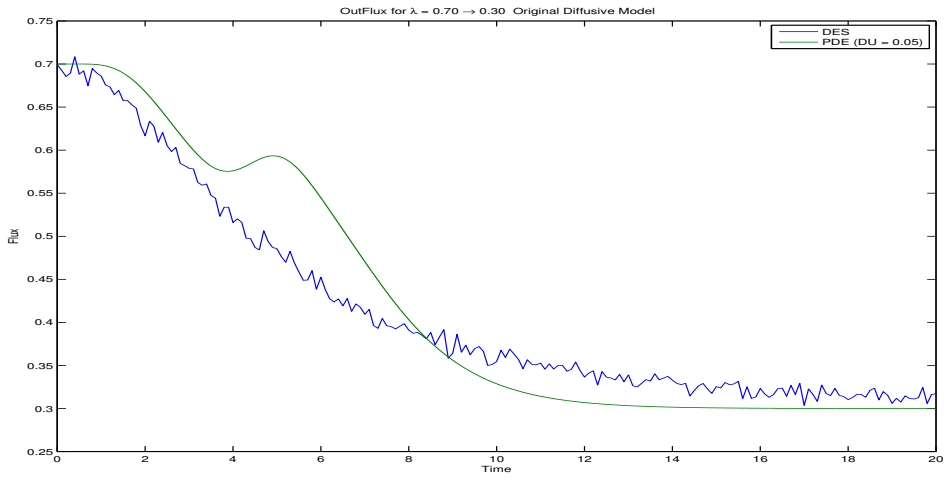
Adding a diffusion term of the form $C\nu_{xx}$, $C \in \mathbf{R}^+$, into Equation (5.16) opposes the backward flow by increasing the velocity of mass that has diffused backwards. Since the velocity defined on the left boundary is inversely proportional to the WIP, this results in a progressively larger gradient in the positive x direction as time goes on and more mass leaves the system. This also means that, so long as C is not too large, this velocity compensation will not significantly affect the corrections that the incorporation of diffusion in (5.15) has made. Rather, the primary effect will be felt as a sizable portion of the initial mass has exited the system and the corrected velocities have finally caught up to the diffused mass. Figure (6.4) suggests that this approach has some merit. Figure (6.4) compares the outflux in time for the original main model (green) and this new model (5.15) with the new velocity equation (red)

$$\nu_t(x,t) + \nu(x,t)\nu_x(x,t) = C\nu_{xx}(x,t). \tag{6.1}$$

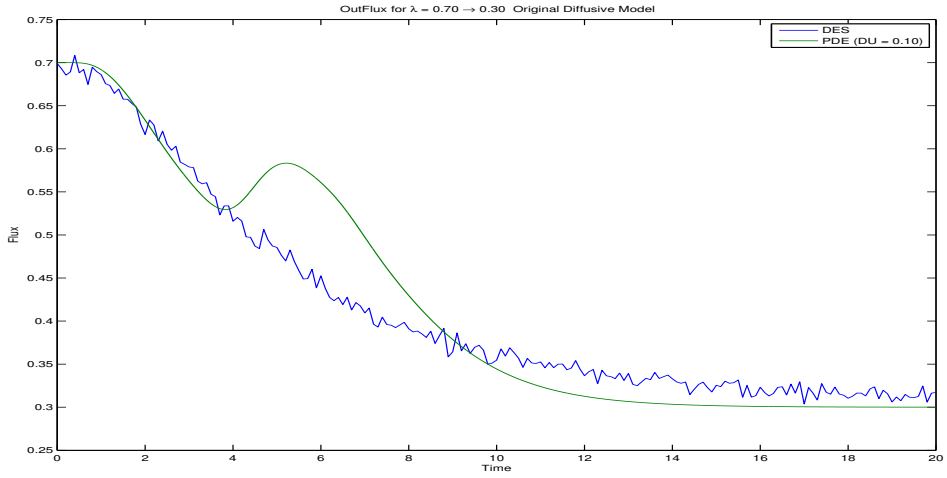
and boundary condition

$$\nu_x(0,t) = 0. \tag{6.2}$$

The previous model only included diffusion in (5.15). With $D = 0.14$ and $C = 0.05$, the backwards flow has been eliminated. In addition, returning to Figure (5.14), one



(a) Diffusion Coefficient 0.05.



(b) Diffusion Coefficient 010.

Figure 6.3: Ramp-Down Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$ with Diffusion.

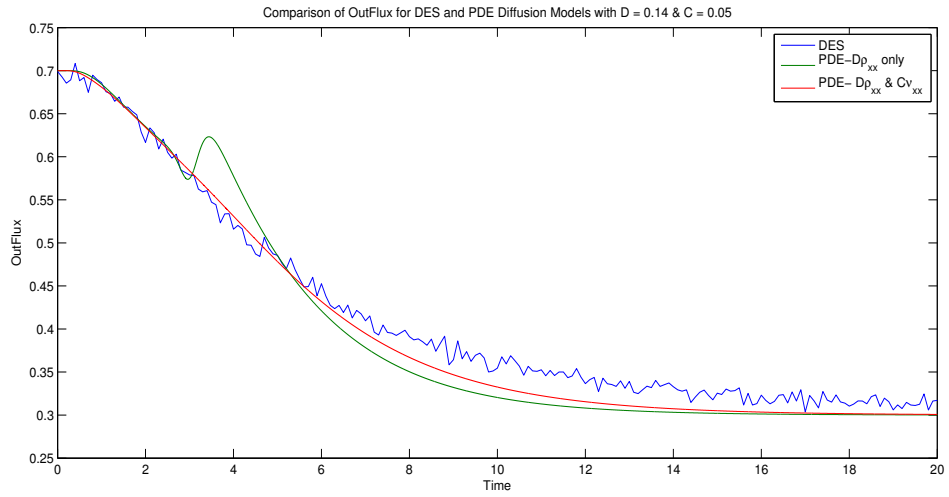
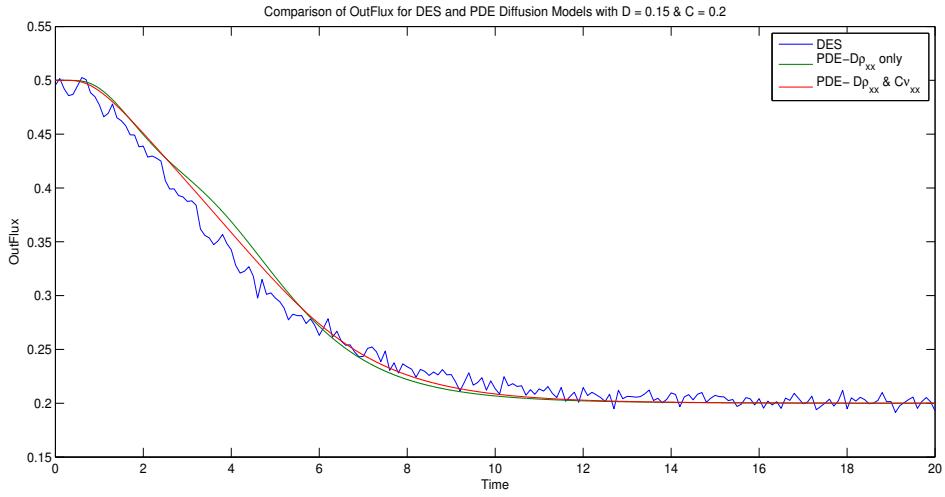


Figure 6.4: Comparison of the Density Diffusion PDE Model with (red) and without (green) Velocity Diffusion for the Ramp-Down Transition $\lambda_1 = 0.7 \rightarrow \lambda_2 = 0.3$.

can see what appears to be nascent backwards flow for the transition $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.2$. Figure (6.5) compares the outflux in time for the two models (5.15) and (6.1). As one can see from the figures, the introduction of a diffusion term to the velocity PDE counters the undesirable effects of the density diffusion without changing the effects of the density diffusion on the initial outflux delay in a significant manner.

While the addition of a diffusion term was due to its mechanism of action on the PDE, its inclusion can be justified heuristically. In the model, the left boundary condition is (5.7). This expression is the mean steady-state velocity for the $M/M/1$ queueing system. For a transient $M/M/1$ queue, this definition of the boundary condition for (5.16) is likely not accurate. This means that when the system evolves there will be additional error incurred that is due solely to the inaccuracy of the model and it is reasonable to assume that this error will grow in time. Adding diffusion is a way of compensating for this modeling error.



0

Figure 6.5: Comparison of the Density Diffusion PDE Model with (red) and without (green) Velocity Diffusion for the Ramp-Down Transition $\lambda_1 = 0.5 \rightarrow \lambda_2 = 0.2$.

REFERENCES

Agnew, C. E., "Dynamic modeling and control of congestion-prone systems", Operations research **24**, 3, 400–419 (1976).

Albey, E., U. Bilge and R. Uzsoy, "An exploratory study of disaggregated clearing functions for multiple product single machine production environments", EP Fitts Department of Industrial and Systems Engineering (2011).

Armbruster, D., P. Degond and C. Ringhofer, "Kinetic and fluid models for supply chains supporting policy attributes", Bulletin of the Institute of Mathematics **66**, 3, 896–920 (2006a).

Armbruster, D., P. Degond and C. Ringhofer, "A model for the dynamics of large queuing networks and supply chains", SIAM Journal on Applied Mathematics **66**, 3, 896–920 (2006b).

Armbruster, D., J. Fonteijn and M. Wienke, "Modeling production planning and transient clearing functions", in "Robust Manufacturing Control", pp. 77–88 (Springer, 2013).

Armbruster, D., D. Marthaler and C. Ringhofer, "Kinetic and fluid model hierarchies for supply chains", Multiscale Modeling & Simulation **2**, 1, 43–61 (2003).

Armbruster, D., D. E. Marthaler, C. Ringhofer, K. Kempf and T.-C. Jo, "A continuum model for a re-entrant factory", Operations Research **54**, 5, 933–950 (2006c).

Armbruster, D. and C. Ringhofer, "Thermalized kinetic and fluid models for reentrant supply chains", Multiscale Modeling & Simulation **3**, 4, 782–800 (2005).

Arrow, K. J., *Studies in the Mathematical Theory of Inventory and Production*, no. 1 (Stanford University Press, Palo Alto, CA, 1958).

Asmundsson, J., R. L. Rardin, C. H. Turkseven and R. Uzsoy, "Production planning models with resources subject to congestion", Naval Research Logistics (NRL) **56**, 2, 142–157 (2009).

Asmundsson, J., R. L. Rardin and R. Uzsoy, "Tractable nonlinear production planning models for semiconductor wafer fabrication facilities", Semiconductor Manufacturing, IEEE Transactions on **19**, 1, 95–111 (2006).

Bramson, M., *Stability of queueing networks* (Springer-Verlag, New York, 2008).

Cercignani, C., *The Boltzmann equation* (Springer-Verlag, 1988).

Dai, J., "On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models", The Annals of Applied Probability pp. 49–77 (1995).

Dai, J., J. Hasenbein and J. V. Vate, "Stability and instability of a two-station queueing network", Annals of Applied Probability pp. 326–377 (2004).

Elmaghraby, S. E., "Production capacity: Its bases, functions and measurement", in "Planning Production and Inventories in the Extended Enterprise: A Sate of the Art Handbook", pp. 119–166 (Springer, New York, 2011).

Eppen, G. D. and R. K. Martin, "Determining safety stock in the presence of stochastic lead time and demand", Management Science **34**, 11, 1380–1390 (1988).

Graves, S. C., "A tactical planning model for a job shop", Operations Research **34**, 4, 522–533 (1986).

Gross, D., J. F. Shortle, J. M. Thompson and C. M. Harris, *Fundamentals of Queueing Theory, Fourth Edition* (John Wiley & Sons, Inc., New Jersey, 2008).

Hackman, S., *Production economics* (Springer-Verlag, 2008).

Hackman, S. T. and R. C. Leachman, "A general framework for modeling production", Management Science **35**, 4, 478–495 (1989).

Haeussler, S. and H. Missbauer, "Empirical validation of meta-models of work centres in order release planning", International Journal of Production Economics **149**, 102–116 (2014).

Harris, F. W., *Operations and cost (Factory management series)* (Shaw, A.W., Chicago, 1915).

Holt, C., *Planning Production, Inventories, and Work Force.* (Prentice-Hall, New Jersey, 1960).

Hung, Y.-F. and R. C. Leachman, "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations", Semiconductor Manufacturing, IEEE Transactions on **9**, 2, 257–269 (1996).

Johnson, L. A. and D. C. Montgomery, *Operations research in production planning, scheduling, and inventory control* (Wiley, New York, 1974).

Kacar, N. B., *Fitting clearing functions to empirical data: Simulation optimization and heuristic algorithms* (North Carolina State University, 2012).

Kacar, N. B. and R. Uzsoy, "Estimating clearing functions from simulation data", in "Proceedings of the Winter Simulation Conference", pp. 1699–1710 (Winter Simulation Conference, 2010).

Karmarkar, U. S., "Capacity loading and release planning with work-in-progress (wip) and leadtimes", Journal of Manufacturing and Operations Management **2**, 105-123 (1989).

Kelton, W. D. and A. M. Law, *Simulation modeling and analysis* (McGraw Hill Boston, 2000).

Kempf, K. G., P. Keskinocak and R. Uzsoy, "Planning production and inventories in the extended enterprise", Springer (2011).

Kim, B. and S. Kim, "Extended model for a hybrid production planning approach", International Journal of Production Economics **73**, 2, 165–173 (2001).

Lefeber, E. and D. Armbruster, "Aggregate modeling of manufacturing systems", in "Planning Production and Inventories in the Extended Enterprise", pp. 509–536 (Springer, 2011).

LeVeque, R. J., *Numerical Methods for Conservation Laws, Second Edition* (Birkhauser Verlag, Basel, Switzerland, 1992).

Levin, D. A., Y. Peres and E. L. Wilmer, *Markov chains and mixing times* (American Mathematical Soc., 2009).

Lewis, P. A. and G. S. Shedler, "Simulation of nonhomogeneous poisson processes by thinning", Tech. rep., DTIC Document (1978).

Medhi, J., *Stochastic models in queueing theory* (Academic Press, New York, 2002).

Missbauer, H., "Aggregate order release planning for time-varying demand", International Journal of Production Research **40**, 3, 699–718 (2002).

Missbauer, H., "Models of the transient behaviour of production units to optimize the aggregate material flow", International Journal of Production Economics **118**, 2, 387–397 (2009).

Missbauer, H., "Order release planning with clearing functions: a queueing-theoretical analysis of the clearing function concept", International Journal of Production Economics **131**, 1, 399–406 (2011).

Missbauer, H. and R. Uzsoy, "Optimization models of production planning problems", in "Planning Production and Inventories in the Extended Enterprise", pp. 437–507 (Springer, 2011).

Modigliani, F. and F. E. Hohn, "Production planning over time and the nature of the expectation and planning horizon", Econometrica, Journal of the Econometric Society pp. 46–66 (1955).

Perdaen, D., D. Armbruster, K. Kempf and E. Lefeber, "Controlling a re-entrant manufacturing line via the push–pull point", International Journal of Production Research **46**, 16, 4521–4536 (2008).

Ross, S. M., *Simulation, Fifth Edition* (Academic Press, Oxford, 2013).

Schneeweiss, C., *Distributed decision making* (Springer, 2003).

Selcuk, B., J. C. Fransoo and A. G. De Kok, "Work-in-process clearing in supply chain operations planning", IIE Transactions **40**, 3, 206–220 (2008).

Spall, J. C., "Implementation of the simultaneous perturbation algorithm for stochastic optimization", Aerospace and Electronic Systems, IEEE Transactions on **34**, 3, 817–823 (1998).

Spearman, M. L., "An analytic congestion model for closed production systems with ifr processing times", Management Science **37**, 8, 1015–1029 (1991).

Srinivasan, A., M. Carey and T. E. Morton, "Resource pricing and aggregate scheduling in manufacturing systems", Tech. rep., Carnegie Mellon University, Tepper School of Business (1988).

Vollmann, T. E., W. L. Berry, D. C. Whybark and F. R. Jacobs, *Manufacturing planning and control for supply chain management* (McGraw-Hill/Irwin New York, 2005).

Zäpfel, G. and H. Missbauer, "Production planning and control (ppc) systems including load-oriented order release problems and research perspectives", International Journal of Production Economics **30**, 107–122 (1993).

Zipkin, P. H., *Foundations of inventory management*, vol. 2 (McGraw-Hill, New York, 2000).